Using Metadata to Automate Interpretations of Unsupervised Learning-Derived Clusters

Richard G. Freedman and Shlomo Zilberstein
University of Massachusetts Amherst
College of Information and Computer Sciences
{freedman, shlomo}@cs.umass.edu

Abstract

Unsupervised machine learning methods are useful for identifying clusters of similar inputs with respect to some criteria and giving the inputs within each cluster the same label. However, the results of many such methods rely on parameter choices that can alter the derived classification labels for each input. Verification methods for determining the quality of clusters often relies on human intuition, but this is not always an easy task depending on the format of the inputs and finding the correct relationship that the algorithm used.We present an approach to assist human verification of the unsupervised learning algorithms' classification choices through the use of metadata describing the inputs that will be clustered. When the metadata measures the relevance of each input to human-interpretable features, we show how a similar measurement of relevance to human-interpretable features can be derived to describe the unsupervised learning algorithm's choices of clusters. An example demonstrating how it can evaluate previous work with activity recognition via topic models is provided in addition to propositions of other uses for the metadata.

1 Introduction

As massive datasets of information become more readily available, there is also a difficulty in properly annotating all the data. Crowdsourcing and domain-specific applications can yield definitive outputs and produce these datasets for supervised machine learning methods with large degrees of accuracy, but other forms of data such as collections of documents and sensor readings are not so easy to analyze. This is one benefit of unsupervised machine learning algorithms that can cluster data without annotation under some sets of parameters. However, the resulting clusters are not always intuitive to a human due to the formulaic procedure of reducing distances or entropy amongst sets of configurations. Even methods such as topic modeling, producing clusters of human-understandable words through mixture and admixture models, do not always generate topic clusters that yield the same interpretation to every person. Research has been done to identify phrases of words within each cluster that can best summarize them compared to the more loosely defined bags of words [Hannah and Wallach, 2014].

Topic models such as latent Dirichlet allocation (LDA) [Blei *et al.*, 2003] have been used for other tasks including activity recognition [Huỳnh *et al.*, 2008], image classification and annotation [Wang *et al.*, 2009], and key-profiling music by its notes [Hu and Saul, 2009]; these new domains derive clusters over other objects rather than word clusters. The primary challenge with these new data formats is the inability for humans to clearly interpret them, leading to difficulty in verification and determining what to do with each cluster. The original work by Huỳnh et al. [2008] provided interpretive evidence for recognizing clusters of wearable sensor data as daily activities by aligning the learned topic clusters with an annotated timeline of activities, but few other applications have had access to such annotations. Furthermore, it is evident that other forms of real-world data can be difficult to display since numbers and configurations are not always easy to relate to one another within a single cluster. Freedman *et al.* [2014] represented activity clusters learned using LDA on red, green, blue, depth (RGB-D) sensor data as collections of stick figures in an attempt to resemble the topic modeling literature where collections of words are presented for each topic. As snapshots of an activity in progress, stick figures and other forms of data visualization still cannot reveal the underlying trend(s) between each other like actual words can because *words have official semantic definitions*.

We thus propose the use of human-interpretable features as metadata, data that describe other data, for unsupervised machine learning algorithm inputs in order to autonomously derive descriptions of learned clusters. This will remove ambiguity in the learned models because the machine can explain the trends in terms that humans comprehend. Metadata has previously been used to describe datasets to assist machine learning algorithms [Cunningham, 1996]. Examples include describing features of datasets in order to determine which supervised machine learning algorithms are most suitable for a new dataset [Brazdil *et al.*, 1994] and providing bibliographical information for individual text documents within a corpus to stratify topics for specific subsets of documents [Mimno and McCallum, 2008]. However, it does not seem that anyone has previously used *metadata to describe the individual data entries*, perhaps due to the large number of unique inputs that can exist within a massive dataset. For each domain, we argue that it should be possible to automate the generation of metadata for each possible input with reasonable computational overhead to avoid this issue.

After defining our metadata representations and providing an example of deriving features for RGB-D sensor data and words in text data in Section 2, we will use them to derive descriptions of learned clusters in their respective domains. Following these formulations, we introduce experiments in Section 3 to explore these labels' usefulness with respect to various factors. Section 4 concludes with a discussion and future work, including how this work may be used to develop more robust unsupervised learning algorithms that can handle on-line data, novel inputs that typically require reclustering or assumptions to classify, and aligning clusters from different runs due to anomolous changes from the value of the random seed.

2 Labeling Metadata for Inputs and Clusters

To derive commonalities between objects in a cluster, we must have a list of human-interpretable properties for each possible object that could be in the input set V for our unsupervised learning algorithm. We define a ***feature vector*** for input $v \in V$ as $\overrightarrow{x_v} \in [0,1]^{|F|}$ where F is the list of possible features and $x_v(i \in F) \rightarrow 1$ as the i^{th} feature is more relevant to v. For describing a particular cluster k's features, we define a ***feature descriptor*** as vector $\overrightarrow{x_k} \in [0,1]^{|F|}$ where $x_k(i \in F) \rightarrow 1$ as the i^{th} feature is more commonly associated with the cluster's inputs and $x_k(i \in F) \rightarrow 0$ as it is less commonly associated.

2.1 Generating Feature Vectors

When processing data to generate the set of inputs for an unsupervised learning algorithm, we propose simultaneously generating each feature descriptor. While the approach for each domain may vary, the measurement of relevance for each feature descriptor's definition should take the form of a checklist that is ideally computationally linear to the number of features $O(|F|)$. We propose examples below for generating this metadata to describe stick figure postures derived from RGB-D sensors when performing wordification [Perovšek *et al.*, 2013] for use in topic models as described by Freedman *et al.* [2015].

Example 1: RGB-D Sensor Data

RGB-D sensor data, collected by such devices as the Kinect, produce a sequence of three-dimensional point clouds that represent a colored surface of the region facing the sensor over time. Each point cloud may be used in activity recognition to represent the environment where regions of changing points over time indicate objects of interest [Zhang and Parker, 2011], and human bodies may be identified from these regions [Shotton *et al.*, 2011] to extract postures independent of the environment [Freedman *et al.*, 2014]. When a person looks at a single posture, she is usually able to explain it in terms of the appendages and joints' relative positions. For example, Fig. 1 is standing with the arms slightly bent, one of which is raised, and one lifted leg that is bent. The conditions for discerning these features are not arbitrary because *specific angles of orientation for each joint dictate the direction and position of the limbs*. As most software packages provide RGB-D sensor data in the form of $[-\pi, \pi]^{45}$ (roll, pitch, and yaw for 15 joints), it is possible to compute Euler angles and determine these features using a list of conditional statements.

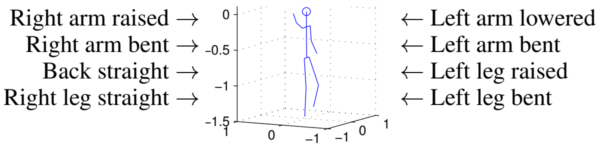

Figure 1: Some human-interpretable features for a posture.

For example, an elbow joint may be considered bent if the angle between the upper and lower arm is in $[0, 3\pi/4]$ and straight if it is in $(3\pi/4, \pi]$. Given lengths for each body link, the Euler angles from the shoulder to the elbow, and the Euler angles from the elbow to the wrist, we can assume that the shoulder is at coordinate location $\vec{s} = [0,0,0]^T$ and then compute the positions of the elbow $\vec{e}$ and wrist $\vec{w}$ with the homogeneous translation and rotation transformations. With these coordinates, the law of cosines may be used to find the angle between the upper and lower arm:

$$\cos(\text{elbow}) = \frac{-|\vec{s}-\vec{w}|^2 + |\vec{e}-\vec{s}|^2 + |\vec{w}-\vec{e}|^2}{2 \cdot |\vec{e}-\vec{s}| \cdot |\vec{w}-\vec{e}|}$$

When the Euler angles are discretized with granularity parameter g to create a collection of inputs with duplicates, the feature can still be *evaluated with a degree of relevance rather than a binary evaluation*. In particular, the discretization yields an interval for each Euler angle, and the ratio of these possible assignments A that correspond to each feature can be computed. For example, the relevance of a bent elbow is

$$x_v(\text{elbow bent}) = \frac{|\{a \in A \,|\text{elbow}_a \in [0, 3\pi/4]\}|}{|A|}$$

where elbow_a is the computed elbow angle with Euler angle assignment a. Measuring relevance is a generalization of the binary approach because a single Euler angle means that $|A| = 1$. Furthermore, when A is infinite (i.e. an interval over real numbers), we can approximate the relevance by sampling random assignments and counting those that satisfy the specific feature. The precise method of identifying boundaries (partitioning A) for each feature may be possible when the number of Euler angles to assign is small, but it may otherwise become a collection of optimization problems that are both concave and convex due to the trigonometric functions involved in homogeneous rotation transformations.

Example 2: Text Document Word Data

Corpora of text data are available everywhere and have been the subject of many research studies ranging from, but not limited to, natural language processing to social science research. Because the underlying concepts in text data are applicable to many of these studies, analyses using latent variables, including the aforementioned topic models, became the standard approach. The term for these types of methods is Latent Semantic Analysis (LSA) [Deerwester *et al.*, 1990] due to the inference of the latent variables' labels. In a discussion similar to the one we present in the introduction, Gabrilovich and Markovitch note that LSA uses statistical means to derive its labels and does not regard the actual semantics that a human understands. They use this argument to derive a new approach titled Explicit Semantic Analysis (ESA) [2009] that takes advantage of large amounts of text data that may be used for creating annotations of other text data.

Specifically, Wikipedia is a collection of text documents where each document provides information about a single word/phrase regarding most, if not all, its concepts and possible interpretations. Similar to how a human can read a dictionary or encyclopedia and use the description to better understand associations between the queried word/phrase and more familiar concepts, Gabrilovich and Markovitch create a sparse matrix of term frequency-inverse document frequency (TFIDF) values for individual terms over each document's primary concept. That is, for term $v \in V$ in document $d_{f \in F}$, the TFIDF value is:

$$M\left[v, f\right] = C_f^{-1} \cdot tf\left(v, d_f\right) \cdot \log \frac{|F|}{|\{i \in F\, |v \in d_i\}|}$$

where $tf\left(v, d_f\right) = 0$ if $v \notin d_f$ and $tf\left(v, d_f\right) = 1 + \log count\left(v, d_f\right)$ if $v \in d_f$, and C_f is the cosine normalization constant over the terms $C_f = \sqrt{\sum_{v \in V} tf\left(v, d_f\right) \cdot \left(\log|F| - \log|\{i \in F\, |v \in d_i\}|\right)}$. Additional modifications based on hyperlink information and generalization filters are applied to remove noise and improve each value's relevance, but we refer the reader to their paper [Gabrilovich and Markovitch, 2009] for more details due to limited space.

The TFIDF values are sufficient to show that the matrix satisfies our definitions as a set of feature vectors for word data found in text documents. Each term's row is a single feature vector where the features $M\left[v \in V, f \in F\right] = \overrightarrow{x_v}\left(f\right)$ are the individual concepts associated with each Wikipedia document. The matrix has been used primarily for identifying features of sentences by adding the feature vectors for each term appearing in the sentence, and we propose extensions of their approach in Section 2.2.

2.2 Generating Feature Descriptors

After learning, we will have K clusters that partition the inputs from the training data. In the case of topic models such as LDA, our inputs are initially grouped into sequences called documents $d \in \{1, \ldots, D\}$, and each input $\overrightarrow{w_{d \in D}}\left(n\right)$ is a word/object. For these sequences, we learn a topic (cluster) assignment $\overrightarrow{z_{d \in D}}\left(n\right)$ for each input. Then each sequence has a distribution of clusters θ_d based on the ratio of cluster labels in $\overrightarrow{z_d}$:

$$\theta_{d \in D}\left(k \in K\right) = \frac{\sum_{n=1}^{|\overrightarrow{z_d}|} \mathbf{1}\left(z_d\left(n\right) = k\right)}{|\overrightarrow{z_d}|}$$

and each cluster k has a distribution over inputs ϕ_k based on the ratio of each input v assigned label k:

$$\phi_{k \in K}\left(v \in V\right) = \frac{\sum_{d=1}^{D} \sum_{n=1}^{|\overrightarrow{w_d}|} \mathbf{1}\left(w_d\left(n\right) = v \wedge z_d\left(n\right) = k\right)}{\sum_{d=1}^{D} \sum_{n=1}^{|\overrightarrow{w_d}|} \mathbf{1}\left(z_d\left(n\right) = k\right)}$$

where $\mathbf{1}$ is the indicator function that equals 1 when the condition is true and 0 otherwise. Smoothing is usually applied based on some hyperparameter settings as well. For other unsupervised learning algorithms with a different formulation, the entire dataset may be a single document ($D = 1$) and duplicate inputs will yield non-uniform distributions for each ϕ_k. The latter condition assumes that the training data is a representative sample of the population so that duplicate inputs indicate a more common object in the population.

Because each θ_d is easily interpreted as a mixture of clusters, we are most interested in finding interpretations for each ϕ_k because the relationships between inputs are not often as obvious. From the activity recognition perspective, we want to identify which features best describe the majority of the sensor readings represented by each cluster's learned distribution. From the text perspective, we want to identify which features define the words in each cluster's learned distribution. Using feature vectors for each input, we propose three approaches for computing feature descriptors for each cluster that can assist with this task:

Expected Value

Due to our definition of a feature vector, each input v assigned to cluster k is located at some point within the $|F|$-dimensional simplex. Because v also has probability mass $\phi_k\left(v\right)$, we can describe the most relevant features of the most common inputs in k as the expected value of each feature $f \in F$: $\overrightarrow{x_k} = \sum_{v \in V} \phi_k\left(v\right) \cdot \overrightarrow{x_v}$. This method is most similar to ESA [Gabrilovich and Markovitch, 2009] because a sentence is described as the sum of the feature vectors for each of its words, which is similar to a distribution over the set of word inputs that is proportional to the word frequencies in the sentence. Although simple to compute, this approach is naïve because it simply finds the weighted union of features. Thus a single v with a large $\phi_k\left(v\right)$ would contribute all its features to the cluster's feature descriptor even if no other objects with considerable mass share some of them.

Agglomerative Clustering

As an alternative to the union of features found in the expected value approach, we also propose a method that includes the intersection of features. Agglomerative clustering hierarchically builds a partition of V such that each subset's inputs that share like features, beginning with singleton subsets that contain each input separately and then iteratively combining similar subsets until the larger paritions are too different to combine. The likeness between two subsets $C_1, C_2 \subseteq V$ with respect to cluster k is measured using

$$d\left(C_1, C_2\right) = \left|\sum_{v \in C_1} \phi_k\left(v\right) - \sum_{v \in C_2} \phi_k\left(v\right)\right| \cdot ||\overrightarrow{x_{C_1}} - \overrightarrow{x_{C_2}}||_1$$

where $\overrightarrow{x_{C_i}}$ is the feature descriptor for subset C_i. d is not a metric because a distance of 0 does not guarantee that the two subsets are equal. However, it does emphasize which *pairs of subsets would have a smaller degree of change when their individual feature descriptors are intersected.* The comparison of probability mass within ϕ_k is also used to avoid placing inputs with lesser representation of cluster k into the same partitions as inputs with greater representation of cluster k.

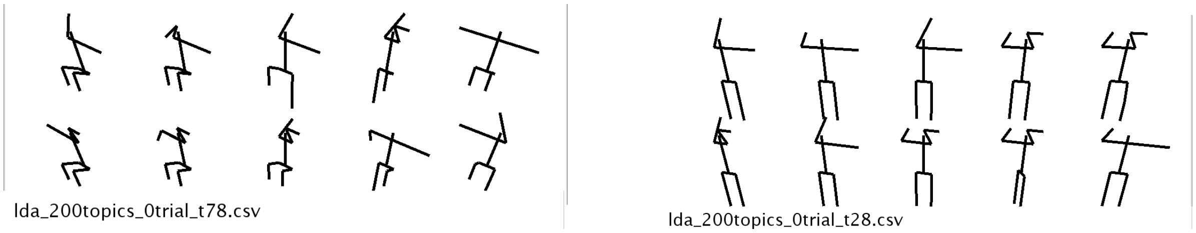

Figure 2: Images of most likely postures for two topics

In contrast to the union of feature vectors, which resembles an expected value, we define the intersection of feature vectors for elements of combined clusters C_1 and C_2 as $\overrightarrow{x_{C_{1,2}}} = \bigodot_{v \in C_1 \cup C_2} \overrightarrow{x_v}$ where $\odot$ is element-wise multiplication. This is consequently the feature descriptor for combined cluster $C_{1,2}$. Intersection may be too strong since it has the opposite problem of the union: a single input with large probability density may lack one feature ($\overrightarrow{x_v}(f) \approx 0$ for some $v \in V$ and $f \in F$) that is greatly relevant to the remaining inputs of significant probability; this feature would be excluded from the cluster's feature descriptor. To address this, we introduce the unweighted average as a ***soft intersection*** that accounts for the number of objects sharing the presence/lack of a feature. We compute $\overrightarrow{x_{C_{1,2}}} = |C_1 \cup C_2|^{-1} \cdot \sum_{v \in C_1 \cup C_2} \overrightarrow{x_v}$ as the soft intersection of the feature vectors of the elements of combined subsets C_1 and C_2. With respect to interpretability, 0 means that a feature is not relevant to any inputs representing the cluster, 1 means that a feature is relevant to all inputs representing the cluster, and a value of 0.5 means that a feature is not useful for a description since it is equally present and absent from the inputs representing the cluster. This appears to have some relevance to the interpretation of entropy, but we do not pursue this connection in this work.

When the distances between subsets becomes too great so that there are no more subsets to combine, then we will have partitions expressing unique features that each describe the cluster; we hypothesize that this will consist of two cases:

- A cluster is equally represented by subsets of distinct features (that is, for each input representing a cluster, there exists at least one subset, but not necessarily all subsets, that describes it) - then $d(C_1, C_2)$ will be too great due to the $\left|\left|\overrightarrow{x_{C_1}} - \overrightarrow{x_{C_2}}\right|\right|_1$ component
- A cluster is composed of a primary subset of features, but some noise during training added unrelated inputs to the cluster - then $d(C_1, C_2)$ will be too great due to the $\left|\sum_{v \in C_1} \phi_k(v) - \sum_{v \in C_2} \phi_k(v)\right|$ component

For either case, we hypothesize that the expected value of the partitions' feature descriptors will be more informative for describing the cluster than the expected value of each object's feature descriptors without any structure. However, there are also advantages to using a disjunction of the subsets' weighted feature descriptors to describe the cluster: an exclusive-or relationship between features may be obscured by combining them. For example, using the RGB-D case, if one set contains postures with "the left arm raised and the right arm not raised" and the other set contains postures with "the left arm not raised and the right arm raised," then this may imply that the cluster contains postures with exactly one arm raised — adding these together for an expected value would instead yield a compromise that the right and left arms may or may not be raised (an irrelevant feature near 0.5).

Supervised Learning

The last approach acknowledges the fact that supervised learning methods such as decision trees learn interpretable functions. For example, the traversal from a decision tree's root to any leaf node produces a conjunction of conditions that explains the leaf's classification assignment. If we consider every input, including duplicates, as a separate data point, then we have supervised inputs $\overrightarrow{x_{w_d(n)}}$ with assigned outputs $z_d(n)$ from our unsupervised learning model. We may use off-the-shelf supervised learning algorithms to learn a function mapping between each feature vector and its associated cluster rather than independently computing feature descriptors for each cluster. Changuel and Labroche [2012] used such off-the-shelf classifiers to learn missing metadata values from present ones to improve categorization of library resources. The only limitation is that each algorithm has a specific type of function which it can learn. For example, decision trees can only learn perpendicular partitions of the feature space. Thus different supervised learning methods may yield different justifications for the unsupervised algorithm's label assignments.

3 Designed Experiments

Previous work by Freedman *et al.* [2014; 2015] used topic models for unsupervised activity recognition by the following analogy between RGB-D sensor data and text documents:

- A document is a single plan execution's recording
- Each frame of the recording's posture is a word
- The topics are activities composing the executed plan, and they represent clusters of postures for the activity

Due to learning the activities without supervision, the only means of verification were those used for validating topics learned from natural language: good log-likehood values for held-out testing sets and viewing the most likely words in each topic. While the log-likehood's interpretation is the same regardless of the data format, it is a relative comparison that indicates "better," but not necessarily "good" or "bad" for each model's representation of the data. Viewing the most likely words in each topic gives humans an opportunity to analyze the cluster on their own, and the authors often provide

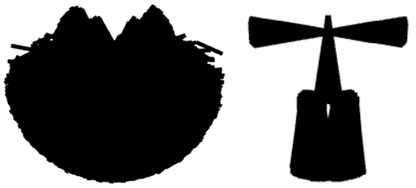

Figure 3: Visualization of 1000000 sampled postures from a single posture discretized with low (3, left) and high (21, right) granularity.

their own expert summary of the displayed topics. However, the most likely postures in each learned activity are not often as obvious to interpret. Figure 2 shows two examples of most likely postures for topics that appear to indicate the activities “sitting” and “one arm outstretched,” but both include similar arm positions. Therefore, *how do we truly evaluate which qualities of the postures represent the activity*? There are thus several experiments that may be applied using feature vectors and feature descriptors to better understand the data and the learned activity recognition model.

3.1 Learned Feature Vectors by Granularity

The first experiment involves one of the earliest points of Freedman *et al.*’s discussion on the knowledge representation of RGB-D data as text: the granularity parameter g. Each joint-angle $\alpha \in [-\pi, \pi]$ constructing the posture could be mapped to an integer i such that $i \leq g \cdot (\alpha + \pi) / (2\pi) < (i+1)$ [Freedman *et al.*, 2014]. Because lesser granularity includes larger intervals of angles per integer, they were avoided so that the visualized postures were restricted for easier visualization. Larger ranges of joint angles allow more possible locations for placing each joint in space, leading to ambiguity as seen in Figure 3 for a posture represented with granularity $g = 3$ when all angles map to $i = 1$. We hypothesize that the feature vector for postures with lower granularity will emphasize this ambiguity during its sampling by computing values of relevance are more uniformly distributed amongst ***complementary features***; that is, mutually exclusive features such as ‘bent limb’ and ‘straight limb’ where $\sum_{f \in G \subseteq F} \overrightarrow{x_v}(f) = 1$ for all $v \in V$. In contrast, feature vectors for lesser granularities should be less uniform and more unimodal between complementary features because, as also shown in Figure 3, there is less ambiguity of the human posture when the joint angles have a smaller interval of possible values.

3.2 Derived Feature Descriptors

Besides being able to interpret inputs individually, it is important to validate that the metadata is useful for making the clusters of inputs more understandable to humans. Even if all the most likely postures can be described autonomously, it is more important that the features they share, and thus what the learned activity represents, are evident. Using the three approaches described in Section 2.2, we intend to investigate the descriptions derived for clusters from Freedman *et al.*’s prior research as well as several natural language text corpora.

Comparisons of the quality of the feature descriptors using the expected value and agglomerative clustering approaches will be important to justify the additional computational resources. Agglomerative clustering seems to be more expressive as it can find disjunctions and conjunctions of shared features, but it requires many distance computations, $O(|V|)$, at each iteration. Furthermore, more memory resources are needed to store feature vectors for the agglomerative clustering approach when subsets of features are complementary. The complementary nature of features such as those in RGB-D posture descriptions allow the representation of feature vectors to be done with some constraints per set of complementary features, even as feature descriptors:

Theorem 1. *Given a set of complementary features $G \subseteq F$, there are $(|G| - 1)$ degrees of freedom for G when computing feature descriptors as linear combinations of feature vectors if the coefficients sum to* 1.

Proof. Let $G \subseteq F$ be a set of complementary features such that $\sum_{f \in G} \overrightarrow{x_v}(f) = 1$ for all $v \in V$. Then let feature $f' \in G$ be the constrained feature:

$$\overrightarrow{x_v}(f') = 1 - \sum_{f \in G \setminus f'} \overrightarrow{x_v}(f).$$

Then the feature descriptor of cluster k formed by a linear combination of feature vectors is $\overrightarrow{x_k}(f \in G) = \sum_{v \in W \subseteq V} \alpha_v \overrightarrow{x_v}(f)$ where $\sum_{v \in W \subseteq V} \alpha_v = 1$. We now consider:

$$\begin{aligned}\overrightarrow{x_k}(f') &= \sum_{v \in W \subseteq V} \alpha_v \left(1 - \sum_{f \in G \setminus f'} \overrightarrow{x_v}(f)\right) \\ &= \sum_{v \in W \subseteq V} \alpha_v - \sum_{v \in W \subseteq V} \alpha_v \sum_{f \in G \setminus f'} \overrightarrow{x_v}(f) \\ &= 1 - \sum_{f \in G \setminus f'} \sum_{v \in W \subseteq V} \alpha_v \overrightarrow{x_v}(f) \\ &= 1 - \sum_{f \in G \setminus f'} \overrightarrow{x_k}(f).\end{aligned}$$

Thus f' is constrained in both the feature vectors and the feature descriptor while the other $|G| - 1$ features are free. □

However, these constrained values are needed when computing distances for agglomerative clustering unless the complementary features all come in pairs. The distances would only be proportional if the complementary features came in pairs — then the distance would be proportional by a factor of 2 because $\left|\left(1 - \sum_{i=1}^{j} a_i\right) - \left(1 - \sum_{i=1}^{j} b_i\right)\right| = \sum_{i=1}^{j} |a_i - b_i|$ when $j = 1$, and other cases are not guaranteed due to triangle inequalities. Such memory considerations are not applicable for feature vectors of text data because the relevant words used as feaures do not have complements. That is, being associated with one conceptual word does not guarantee a disassociation with another conceptual word.

In addition to comparing the trade-offs between computational complexity and quality of feature descriptors, we also need to look into the advantages and disadvantages of feature descriptors in comparison to functions learned by interpretable supervised learning algorithms such as decision trees. The greatest difference between these two approaches for explaining clusters is that feature descriptors are explanations of individual clusters, independent of the other ones. We hypothesize that this will be useful for interpreting what features describe each cluster. However, unsupervised learning algorithms partition the inputs into clusters so that there are relationships between them, and these are not captured by considering the inputs exclusively in a single cluster. We hypothesize that the supervised machine learning methods can provide insight into the *distinguishing features* that make each cluster unique. Learning functions with poor performance (precision, recall, etc.) may even provide insight into whether too many were chosen for the unsupervised learning parameter because inputs that should be in the same cluster may be split into the unnecessary additional clusters. Thus, we want to determine whether the *intercluster comparisons and intracluster features are mutually exclusive or have some overlap of information*.

4 Discussion

Unsupervised learning algorithms have been useful for autonomously assigning labels when there is data that is difficult to manually label either due to the amount of necessary manpower or due to the challenge of selecting the correct label. However, this convenience comes at the price of interpretability because the optimization algorithms used to cluster inputs into each label do not consider standard patterns that a human would observe. To aid humans in understanding these learned clusters so that they may interpret the labels, we introduced a data structure made of metadata whose features describe the inputs in a human context. We then proposed how to use the feature vectors for a range of tasks including evaluations of discretization choices for continuous input spaces, deriving similar metadata to describe the learned clusters over inputs, and comparing the features between clusters learned in a single training session. In addition to the RGB-D posture and text document domains provided as examples, we believe that domain experts can create expert systems to autonomously generate feature vectors for their respective datasets to produce similar human-interpretable explanations of clusters that allow us to go beyond the label from the classifier.

4.1 Other Potential Applications

The feature descriptors' ability to extract the defining features of a cluster may be used for more than just deriving human-interpretable explanations. It may also be used as a *computational tool for comparisons* when using learned unsupervised models for classification as well as when continuing the learning process with additional training samples. For on-line classification, optimization-based clustering methods such as k-means typically compare distances of the new object v''s features to a specific point (usually the centroid) of each cluster. However, this point may not be the exact center depending on the training data and the distance function compares *all the features in the vectors rather than the ones relevant to the cluster*. Inference-based methods such as topic models present similar classification issues using distributions conditioned on previous cases of observing v'. We hypothesize that computing the distance between v''s feature vector and each cluster's feature descriptor instead compares v' with a generalized set of features for the entire cluster with a focus on the more relevant features.

Due to this, computing sufficiently large distances from each cluster should indicate that v' is novel and does not belong in any of the current clusters. Handling novel inputs has been referred to as the domain adaptation problem [Jiang, 2008] due to the need to address cases during testing for which the training data did not prepare the learned classifier. Some researchers omit this concern when using the classifier as a codebook for the purpose of reducing the cardinality of a large space of objects [Zhang and Parker, 2011; Wang and Mori, 2009], but others rely on nonparametric Bayesian processes such as the Pitman-Yor Process [Pitman and Yor, 1997] to dynamically determine the number of clusters to learn. The latter is more common during training than testing, but our distance method can create a new cluster containing just v' on-line. When the system is not running at a later time, it may recluster in case the new inputs change the composition of other clusters.

This training must often be done incrementally, though. When the unsupervised algorithm relies on random sampling methods, different seeds and permutations of the inputs in the training data will yield different clusters, often permutations of one-another. Referred to as the label-switching problem [Redner and Walker, 1984; Stephens, 2000], aligning these permuted clusters to speed up training through parallel execution is difficult. Many methods have already been proposed to compare the clusters' distributions [Jasra *et al.*, 2005], and we are interested in comparing the resulting matchings and runtimes between these approaches and our proposed application of feature descriptors.

4.2 Future Work

To verify the extent to which our proposed methods can help explain unsupervised learning-derived clusters, we will implement generators for feature vectors for both RGB-D postures and words in text data. These may be used to illustrate how ambiguity of multiple interpretations for a single posture or word can be clarified with weighted feature descriptors. This will include a comparison of the three approaches to determine whether it is worth additional computation overhead (expected value versus agglomerative clustering) and whether comparison between all clusters is better than consideration independently (feature descriptors versus supervised machine learning-derived functions). In addition to the aid of interpretation, we will investigate the other applications proposed for this form of metadata. In particular, it would be ideal that the metadata representation not only improves the understanding of how the unsupervised machine learning algorithm is assigning labels, but also assists in other association tasks that require a better understanding of the labeling process.

Acknowledgments

This work was supported in part by the National Science Foundation grant number IIS-1405550.

References

[Blei *et al.*, 2003] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

[Brazdil *et al.*, 1994] Pavel Brazdil, João Gama, and Bob Henery. Characterizing the applicability of classification algorithms using meta-level learning. In Francesco Bergadano and Luc De Raedt, editors, *Machine Learning: ECML-94*, volume 784 of *Lecture Notes in Computer Science*, pages 83–102. Springer Berlin Heidelberg, 1994.

[Changuel and Labroche, 2012] Sahar Changuel and Nicolas Labroche. Content independent metadata production as a machine learning problem. In Petra Perner, editor, *Machine Learning and Data Mining in Pattern Recognition*, volume 7376 of *Lecture Notes in Computer Science*, pages 306–320. Springer Berlin Heidelberg, 2012.

[Cunningham, 1996] Sally Jo Cunningham. Dataset cataloging metadata for machine learning applications and research. In *Proceedings of the Sixth International Workshop on AI and Statistics*, 1996.

[Deerwester *et al.*, 1990] Scott Deerwester, Susan T. Dumais, George W. Furnas, Thomas K. Landauer, and Richard Harshman. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6):391–407, 1990.

[Freedman *et al.*, 2014] Richard G. Freedman, Hee-Tae Jung, and Shlomo Zilberstein. Plan and activity recognition from a topic modeling perspective. In *Proceedings of the Twenty-Fourth International Conference on Automated Planning and Scheduling*, pages 360–364, Portsmouth, New Hampshire, USA, 2014.

[Freedman *et al.*, 2015] Richard G. Freedman, Hee-Tae Jung, and Shlomo Zilberstein. Temporal and object relations in unsupervised plan and activity recognition. In *Proceedings of AAAI 2015 Fall Symposium on AI for Human-Robot Interaction*, pages 51–59, Arlington, Virginia, USA, 2015.

[Gabrilovich and Markovitch, 2009] Evgeniy Gabrilovich and Shaul Markovitch. Wikipedia-based semantic interpretation for natural language processing. *Journal of Artificial Intelligence Research*, 34(1):443–498, 2009.

[Hannah and Wallach, 2014] Lauren A. Hannah and Hanna M. Wallach. Summarizing topics: From word lists to phrases. In *NIPS 2014 Workshop on Modern Machine Learning and Natural Language Processing*, pages 1–5, Montreal, Quebec, Canada, 2014.

[Hu and Saul, 2009] Diane Hu and Lawrence K. Saul. A probabilistic topic model for unsupervised learning of musical key-profiles. In *Proceedings of the Tenth International Society for Music Information Retrieval Conference*, pages 441–446, Kobe, Japan, October 2009.

[Huỳnh *et al.*, 2008] Tâm Huỳnh, Mario Fritz, and Bernt Schiele. Discovery of activity patterns using topic models. In *Proceedings of the Tenth International Conference on Ubiquitous Computing*, pages 10–19, Seoul, South Korea, 2008.

[Jasra *et al.*, 2005] Ajay Jasra, Chris C. Holmes, and David A. Stephens. Markov chain monte carlo methods and the label switching problem in Bayesian mixture modeling. *Statistical Science*, 20(1):50–67, February 2005.

[Jiang, 2008] Jing Jiang. A literature survey on domain adaptation of statistical classifiers. http://sifaka.cs.uiuc.edu/jiang4/domain_adaptation/survey/da_survey.pdf, 2008.

[Mimno and McCallum, 2008] David Mimno and Andrew McCallum. Topic models conditioned on arbitrary features with dirichlet-multinomial regression. In *Proceedings of the Twenty-Fourth Conference Annual Conference on Uncertainty in Artificial Intelligence*, pages 411–418, Corvallis, Oregon, USA, 2008.

[Perovšek *et al.*, 2013] Matic Perovšek, Anže Vavpetič, Bojan Cestnik, and Nada Lavrač. A wordification approach to relational data mining. In Johannes Fürnkranz, Eyke Hüllermeier, and Tomoyuki Higuchi, editors, *Discovery Science*, volume 8140 of *Lecture Notes in Computer Science*, pages 141–154. Springer Berlin Heidelberg, 2013.

[Pitman and Yor, 1997] Jim Pitman and Marc Yor. The two-parameter Poisson-Dirichlet distribution derived from a stable subordinator. *The Annals of Probability*, 25(2):855–900, April 1997.

[Redner and Walker, 1984] Richard A. Redner and Homer F. Walker. Mixture densities, maximum likelihood and the EM algorithm. *SIAM Review*, 26(2):195–239, 1984.

[Shotton *et al.*, 2011] Jamie Shotton, Andrew Fitzgibbon, Mat Cook, Toby Sharp, Mark Finocchio, Richard Moore, Alex Kipman, and Andrew Blake. Real-time human pose recognition in parts from a single depth image. In *Proceedings of the Twenty-Fourth IEEE Conference on Computer Vision and Pattern Recognition*, pages 1297–1304, Colorado Springs, Colorado, USA, 2011.

[Stephens, 2000] Matthew Stephens. Dealing with label switching in mixture models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 62(4):795–809, 2000.

[Wang and Mori, 2009] Yang Wang and Greg Mori. Human action recognition by semi-latent topic models. *IEEE Transactions on Pattern Analysis and Machine Intelligence Special Issue on Probabilistic Graphical Models in Computer Vision*, 31(10):1762–1774, 2009.

[Wang *et al.*, 2009] Chong Wang, David M. Blei, and Fei-Fei Li. Simultaneous image classification and annotation. In *2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 1903–1910, Miami, Florida, USA, June 2009.

[Zhang and Parker, 2011] Hao Zhang and Lynne E. Parker. 4-dimensional local spatio-temporal features for human activity recognition. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 2044–2049, San Francisco, California, USA, 2011.