Automated Feature Learning: Mining Unstructured Data for Useful Abstractions

Abhishek Bafna
EECS, University of Michigan, Ann Arbor
abafna@umich.edu

Jenna Wiens
EECS, University of Michigan, Ann Arbor
wiensj@umich.edu

***Abstract*—When the amount of training data is limited, the successful application of machine learning techniques typically hinges on the ability to identify useful features or abstractions. Expert knowledge often plays a crucial role in this feature engineering process. However, manual creation of such abstractions can be labor intensive and expensive. In this paper, we propose a feature learning framework that takes advantage of the vast amount of expert knowledge available in unstructured form on the Web. We explore the use of unsupervised learning techniques and non-Euclidean distance measures to automatically incorporate such expert knowledge when building feature representations. We demonstrate the utility of our proposed approach on the task of learning useful abstractions from a list of over two thousand patient medications. Applied to three clinically relevant patient risk stratification tasks, the classifiers built using the learned abstractions outperform several baselines including one based on a manually curated feature space.**

I. Introduction

Informative features are often instrumental for the successful application of machine learning techniques. When the number of training examples is limited, an appropriate feature space can lead to the identification of important similarities among examples. Domain-specific knowledge often plays a key role in engineering such features or *abstractions*. For instance, given a long list of medications, an expert in pharmacology can create a taxonomy for the drugs based on their molecular structure and/or physical and chemical properties. This kind of expert abstraction can encode crucial information necessary for identifying similarities among patients. E.g., such abstractions could help identify two patients, one taking penicillin and another taking amoxicillin, as similar.

Here, we explore novel feature-learning techniques for automatically incorporating auxiliary expert knowledge. We focus on building abstractions from categorical data since such data are common across a number of different domains including healthcare, social sciences, recommender systems, etc. The often high dimensionality of these data can make it difficult to identify relationships using standard statistical methods [1]. In order to obtain a robust result, the amount of data required can grow exponentially with the number of categories [2]. Moreover, traditional Euclidean-based measures of similarity generally fail to capture complex relationships among categories/features. Thus, expert knowledge is typically required to identify these relationships and engineer useful lower-dimensional representations of the data.

Unfortunately, the process of manually designing abstractions using expert knowledge is labor- and time-intensive, in addition to being subjective. This has led researchers to devise more automated methods for building abstractions [3]–[6]. However, when the data are categorical (like in the example described above), conventional unsupervised dimensionality reduction techniques (e.g., Principal Component Analysis (PCA), Latent Semantic Indexing (LSI), Locally Linear Embedding (LLE)) can miss important domain-specific relationships [7]–[9]. We aim to solve this problem through the incorporation of readily available expert knowledge, extracted in an unstructured form from the Web.

In this paper, we present an unsupervised feature-learning framework for building useful abstractions for categorical data. Our framework has two stages:

1) Using unstructured data procured from the Web, we learn a hierarchical Pachinko allocation model (HPAM) [10] to discover a set of latent variables (i.e., supertopics and subtopics) that describe relationships among the categories.
2) Using the Earth Mover's Distance (EMD), we account for the non-uniform distances between the latent variables [11], [12]. This enables us to build a relevant similarity matrix that can be used for both supervised and unsupervised learning tasks.

In the first stage, we employ hierarchical topic models for discovering semantic relations among categories. In doing so, we exploit the innate hierarchical structure often present in categorical data. Applied to data represented by high-dimensional categorical features, this stage results in a lower-dimensional feature representation. In the second stage, we further capture the complex relationships among categories by implicitly incorporating the *ground distance* (i.e., the underlying distance between individual features) when computing similarity between examples.

As a case study, we consider the task of learning useful abstractions from a list of over 2,000 medications. In healthcare applications, patient medications are typically encoded using an n dimensional feature vector, where n is the number of distinct medications [13]–[15]. However, patients are typically on only a few medications at a time. Identifying similar patients in such a sparse, high-dimensional feature space is challenging. Thus, using our proposed framework, we leverage the large amount of pharmacology knowledge available on the Web to learn a more useful feature representation.

We apply this representation to over 25,000 patient admissions and build a non-linear kernel representing patient similarities. We demonstrate the utility of this approach, by

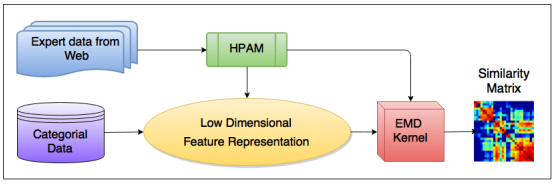


Fig. 1: Our proposed feature-learning framework is composed of two major stages: 1) we learn a low-dimensional representation of the data using a hierarchical topic model applied to unstructured text data and 2) we apply a non-Euclidean distance measure (the EMD) to generate a similarity matrix for the original high-dimensional data.

considering three different clinical prediction tasks. When the amount of training data is limited, our approach outperforms a number of baselines, including a classifier based on a manually curated taxonomy, on all three prediction tasks.

II. Background/Related Work

Typically, in machine learning applications, a categorical variable with n categories is represented by an n dimensional feature vector. When n is large and each example is associated with only a small number of possible categories, the resulting feature space is both high dimensional and sparse. These issues make it difficult to identify similar examples.

These issues are compounded by the fact that such a representation ignores any underlying relationships among categories. In some applications, there exists readily available information regarding such relationships. E.g., in image processing and color perception, the categorical values (colors) can be easily represented by a vector of RGB intensities [16], [17]. This additional information results in a more meaningful comparison than if we compared the categorical values alone. However, in many other applications, quantifying the relationships among categories is not as straightforward. To this end, unsupervised dimensionality reduction techniques like *k-means* clustering have been extended to include categorical information [18]–[20].

Although these techniques aim to discover abstractions from the data itself, in many cases expert or domain-specific knowledge is essential for building informative low-dimensional representations of categorical data. Ontology and taxonomy development have played an important role in the conceptualization of domain-specific knowledge for the purpose of information retrieval and interoperability between domains. However, manually curated ontologies remain laborious, expensive, and challenging to maintain in dynamic domains [21], [22]. This has led to the fast rise of ontology and taxonomy learning systems that automatically or semi-automatically discover knowledge. These systems rely on a variety of data inputs, from structured to unstructured, to generate taxonomies [4], [23]–[29]. To date, the work in this area has been primarily focused on lexical semantics and on applications like information retrieval and question answering.

Here, we consider these kinds of learning systems in a different setting, one that is focused on feature engineering. We induce a taxonomy-structure from unstructured data as a means to engineer informative low-dimensional features and capture important relationships in categorical data. Our goal is to learn a useful representation of the data that leads to the identification of important similarities. We describe our feature-learning framework in the sections that follow.

III. Methods

In this section, we present a formal description of the problem statement and introduce notation. We then describe the different stages of our feature-learning framework.

A. Notation and Problem Statement

Consider a dataset of N data points $\{\mathbf{x}_i|\mathbf{x}_i \in \mathcal{X}\}_{i=1}^{N}$, where $\mathcal{X}$ is an M-dimensional binary feature space, i.e., $\mathbf{x}_i$=$\{x_{i1}, x_{i2}, \ldots, x_{iM}\}$ and $x_{ij} \in \{0, 1\}$.

In addition to these categorical data, we have a corpus of M documents, consisting of unstructured text data. We assume that each of the M documents pertains to a single category in $\mathcal{X}$. Given these data, we aim to learn a low-dimensional feature representation, $\mathbf{x'}_i \in \mathbb{R}^k$ where $k < M$.

B. Hierarchical Pachinko Allocation Model

Given the corpus of unstructured text data, where each document corresponds to a category, we first learn a low-dimensional shared representation of these data (i.e., topics). We begin by modeling the unstructured text data using an HPAM, a variant of the Pachinko allocation model (PAM) [30]. A PAM uses a directed acyclic graph (DAG), where the nodes represent topics, to capture relationships among topics. In a PAM, a subset of topics (i.e., supertopics) is associated with a distribution over other topics (i.e., subtopics) and only these subtopics are associated with a distribution over the vocabulary (i.e., word distribution). An HPAM also captures the hierarchical relationships among topics, but in contrast to a PAM, every topic is associated with a word distribution. Here we consider the second variant of HPAM, presented in [10] and illustrated in Figure 2.

In this model the Dirichlet distribution of each internal node has one extra 'exit' dimension. In our case, the root node has an $R+1$ - dimensional distribution over the supertopics and each supertopic has an $S+1$ - dimensional distribution over the subtopics. This additional dimension corresponds to the event that a word is sampled directly from the word distribution associated with that topic.

While LDA and its variants do capture correlation patterns in words, they are limited in their ability to capture correlations among the topics themselves. But in most real-world data, topics are generally not independent of each other. Ignoring these correlations can lead to the discovery of incoherent topics. Also the categorical data might possess innate hierarchical structure. A PAM captures all correlations between topics using a directed acyclic graph (DAG). Though it represents a nested hierarchy of topics, it limits the association between the distribution of words, from the vocabulary, to only the subtopics. An HPAM, on the other hand extends this association to every node, thus providing more flexibility in the way the words are sampled.

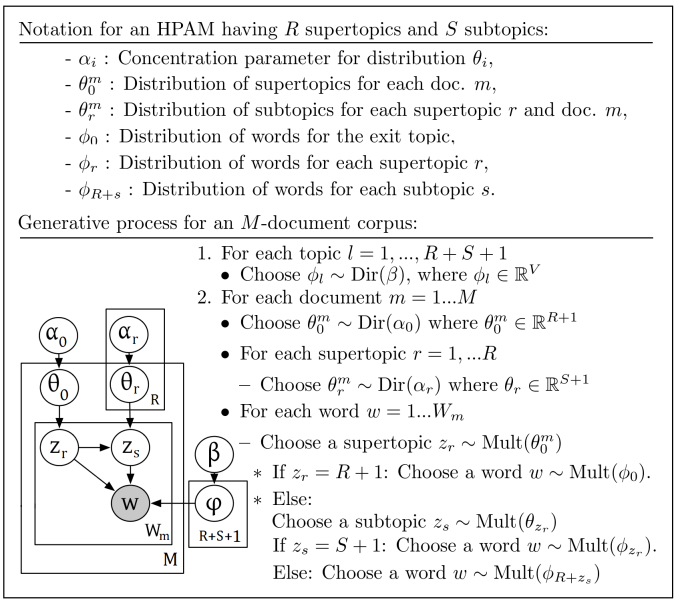


Fig. 2: An HPAM, through a set of supertopics and subtopics, captures the innate hierarchy present within a corpus of documents. θ_0 represents a distribution over supertopics while θ_r represents a distribution over subtopics for supertopic r.

C. Low-Dimensional Representation

Next, we employ the topic model learned in the previous section to construct a low-dimensional representation of the dataset $\{\mathbf{x}_i\}_{i=1}^N$. Given the topic model, we represent each document in our corpus using a feature vector, $\mathbf{z}_m = [\theta_0^m; \Theta_m^T \theta_0^m]$, where $\theta_0^m \in \mathbb{R}^{R+1}$ is its distribution over supertopics, $\theta_r^m \in \mathbb{R}^{S+1}$ for $r = 1 \dots R$, are its distributions over subtopics, and $\Theta_m = [\theta_1^m; \theta_2^m; \dots; \theta_R^m]$.

For the data point i, we have a high-dimensional feature vector $\mathbf{x}_i \in \{0,1\}^M$. Let $|\mathbf{x}_i|$ represent the number of non-zero features. Then we can derive a lower-dimensional representation of $\mathbf{x}_i$, $\mathbf{x'}_i \in \mathbb{R}^k$ where $k = R+S$ (the combined number of supertopics and subtopics) as :

$$\mathbf{x'}_i = \frac{1}{|\mathbf{x}_i|} \sum_{m=1}^{M} \mathbb{1}_{\{x_{im}=1\}} \mathbf{z}_m \tag{1}$$

where $\mathbb{1}_{\{.\}}$ is an indicator function. Note that we do not include the "exit" topics in our final feature vectors since we do not expect these topics to be informative.

D. Measuring Similarity

Given the feature representation obtained in the previous section, we could quantify similarity based on the Euclidean distance in the low-dimensional space. However, this assumes that the data lie in an orthogonal feature space. This assumption does not hold in our application, since some features/topics are closer to each other than to others. Thus we consider a similarity measure based on a cross-bin comparison that takes into account the distance between topics represented by the word distributions. We use the Earth Mover's Distance (EMD) to efficiently incorporate this distance.

The EMD is based on a solution to the well-known transportation problem [31]. It gives a minimal cost transformation from one distribution to another, where cost is determined by the flow between the bins of respective distributions, and the ground distance between them. In our case, the ground distance matrix is determined by computing all pairwise distances between the supertopics and subtopics. These topics are themselves defined as a distribution over the vocabulary (w), and thus we use a symmetric version of the Kullback-Leibler divergence [32] to compute the distance between two topics represented by their word distributions.

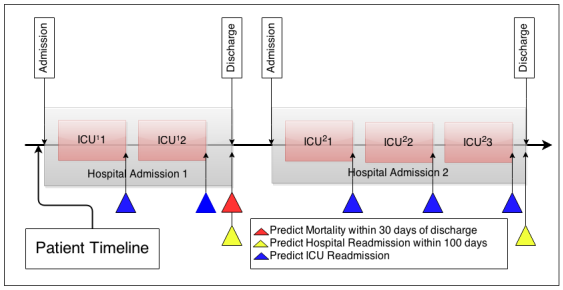


Fig. 3: As a case study, we consider three different prediction tasks. Predictions are made at different time points, depending on the task. As shown, a patient can have multiple hospital admissions and within each admission have multiple ICU visits. This results in multiple predictions per patient.

Given this ground distance, we compute the EMD between each pair of examples $EMD(\mathbf{x}_i, \mathbf{x}_j)$ as in [11]. We then transform this distance matrix into a similarity matrix, $\mathbf{K}$, by replacing the squared-Euclidean distance in the RBF kernel with the EMD:

$$\mathbf{K}(\mathbf{x}_i, \mathbf{x}_j) = exp(-\frac{EMD(\mathbf{x}_i, \mathbf{x}_j)}{2\sigma^2}) \tag{2}$$

where $\mathbf{x} \in \mathbb{R}^k$. $\mathbf{K}$ can be used in a number of learning applications, including both supervised and unsupervised tasks. However, before employing $\mathbf{K}$ as a kernel in kernelized-methods, it is important to note that it is not guaranteed to be positive semi-definite (PSD) [33]. Still, empirically, such kernels have been shown to often work well in practice [34], [35]. In the next section, we describe a series of experiments designed to measure the utility of the resulting similarity matrix and our proposed feature-learning framework.

IV. Case Study

To test the utility of our feature-learning framework we consider a healthcare application in which we focus on representing patients in the intensive care unit (ICU) in terms of their prescribed medications. We compare our approach to a number of other feature representations. We begin by describing the dataset and the classification tasks we consider.

A. Dataset and Classification Tasks

We use the MIMIC II Clinical Database [36]. This publicly available database contains clinical data from approximately 32,000 ICU patient admissions. For the purpose of this study, we focus on representing each patient admission by the pharmacy orders (i.e., ordered medications). Using these data we aim to build classifiers for predicting patient risk for the adverse outcomes described in Table I.

For the first task, we consider all admissions in which the patient is discharged alive from the hospital. We make

our predictions at the time of discharge and consider all medications prescribed during that admission. Patients can have multiple hospital admissions (see Figure 3). Here, we focus on predictions based on the *first* hospital admission; this results in a single prediction per patient. For the second task, we focus on all the admissions of the patient. This results in a single prediction for each admission. Therefore, the number of total examples is higher than in the first task (see Table I). For the third task, we consider all ICU visits in which a patient was discharged alive. Since patients can have multiple ICU visits per hospital admission, we have the greatest number of examples for this task. Similar to the tasks above, when predicting readmission to the ICU we consider only those medications ordered during the most recent ICU visit. For all tasks, we omit visits for which there were no recorded medications.

Outcome	Total #	Positive #
Mortality within 30 days of discharge	22,949	2,828
Readm. within 100 days of discharge	23,292	2,082
Readm. to the ICU during the same visit	23,998	5,587

TABLE I: We consider three clinically relevant classification tasks. All three tasks have a high degree of class imbalance. (Readm. = Readmission)

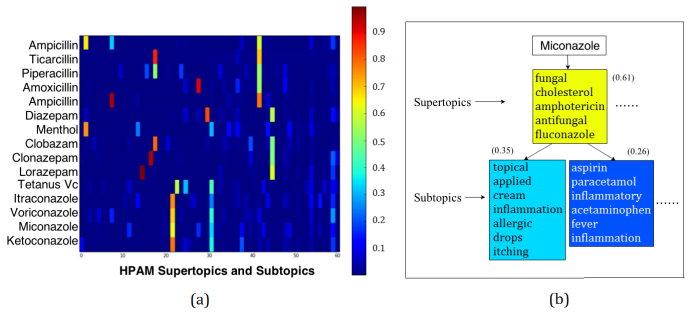
We extract information regarding the medication orders for each example. In the MIMIC II database physician order entries for medications are represented by a free-text entry and a timestamp. Based on the free-text entries for the entire population, we started with a list of 2,285 unique drug names. After removing erroneous entries, 2,164 unique drug names remained. Therefore, each patient may be represented by a feature vector with 2,164 categories. In the next section we apply our proposed feature-learning framework to learn a lower-dimensional representation.

B. Model Construction

1) Web Data Mining: In order to learn a meaningful low-dimensional representation, we required expertise in pharmacology. For this we turned to the Web, using Google and Wikipedia as our auxiliary source of "expert" knowledge. We queried each of the 2,164 drugs using Google and downloaded the corresponding Wikipedia article. (We also derived similar results from specialized web-sites like *Rxlist.com* but the results are not shown here). This resulted in a corpus of 1,124 unique Wikipedia articles, since many of the drugs mapped to the same URL.

Next, we extracted plain text from each Wikipedia page in our corpus. We applied standard preprocessing techniques (e.g., filtered stop words). Our final vocabulary consisted of 21,785 words.

2) Learning the HPAM: From the Wikipedia corpus, we built an HPAM with 30 supertopics and 30 subtopics. We learned the super/subtopic and word distributions using the Gibbs EM algorithm [37], [38]. We determined the total number of topics, $R + S = 60$, using a hierarchical Dirichlet process [39]. We selected the ratio of supertopics to subtopics using five-fold cross validation on the corpus by maximizing the likelihood of the held-out fold. We estimated the other hyperparameters, including the concentration parameters of the Dirichlet priors, using the fixed-point iteration method described in [40].



(a) (b)

Fig. 4: (a) The distribution over learned topics for a subset of drugs. We observe that important relationships among medications are successfully captured by the model. E.g., a number of related antimicrobials (e.g., ampicillin, ticarcillin, etc.) all share the same topic. (b) Miconazole, a topical antifungal medication: the most probable words associated with miconazole's most probable supertopic are given, along with the two most probable subtopics. Probabilities are shown in parentheses.

As an example, in Figure 4, we show the major supertopic and subtopics corresponding to the medication Miconazole. Miconazole comes as a cream, powder or liquid and is an antifungal agent used to treat topical skin infections. We see that these traits are effectively captured by the learned model.

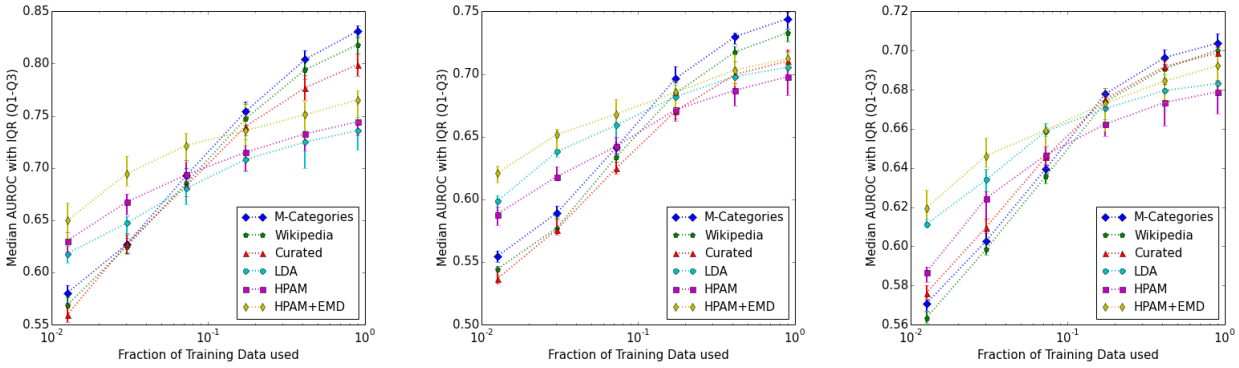
C. Feature Representations

Using the HPAM we transform the high-dimensional categorical patient data into low-dimensional feature vectors as in Section III-C. These low-dimensional features can then be compared using the EMD-based similarity measure from equation 2. Along with the labels, this similarity matrix can be used to learn a classifier. We refer to this approach as *HPAM+EMD* in the next section. Here, we present several additional feature representation approaches. In the next section, we compare the performance of these approaches to our approach.

M-Categories: Each category is simply mapped to its own feature/variable. Applied to our data this approach represents each patient by a feature vector with 2,164 dimensions. On average 0.79% of the vector is nonzero.

Curated: As a proxy for a manually created taxonomy, we chose a taxonomy freely available from the website *Drugs.com*, which manages an extensive library of information including content U.S. Food and Drug Administration (FDA) making it an ideal proxy for our study. Applied to our dataset, we managed to classify all but 92 of the 2,164 drugs into 305 different classes. The remaining 92 medications, which were not explicitly included in the manually curated taxonomy, were included as individual features. This resulted in a feature vector with 397 features.

Wikipedia: When querying each drug name, at times multiple drug names corresponded to the same Wikipedia page. This in itself is a form of abstraction, since features that map to the same Wikipedia page can be combined into one feature. This representation results in a feature vector with 1,124 dimensions, where each dimension corresponds to a unique Wikipedia page.



(a) 30-day Mortality (b) 100-day Hospital Readmission (c) ICU Readmission

Fig. 5: When the amount of training data is limited (e.g., <10%), HPAM+EMD consistently outperforms the other approaches, including one based on a curated taxonomy.

LDA: To measure the utility of choosing HPAM over other topic modeling approaches, we also learn a latent Dirichlet allocation (LDA) model. For the purpose of comparison, we constructed this using the same number of topics as the HPAM, and optimized the Dirichlet hyperparameters using the fixed-point iteration method [40]. Based on the learned topics, we construct patient feature vectors using the approach described in III-C.

HPAM: We also use the abstractions generated from the HPAM as a baseline approach and compare it to our feature representation. This is primarily to test the utility of the EMD distance metric in estimating the distances between the different topics generated by the model.

D. Learning the Classification Models

In this section, we compare the utility of these approaches for learning a classifier for each of the three classification tasks described earlier. For each outcome (i.e., task), we use support vector machines (SVMs) to learn a mapping from each feature representation to patient risk. We train a linear kernel SVM for each of the baseline approaches. (We also considered an RBF kernel, however the results were no better than the linear kernel and are thus not shown here.) In addition, we learn an SVM using our precomputed *HPAM+EMD* kernel.

For training and testing, we generate 50 random stratified splits of 0.7 (training) to 0.3 (test). We set the SVM cost hyperparameter and the bandwidth for the EMD kernel using 5-fold cross-validation on the training data, optimizing for the area under the operating characteristic curve (AUROC). We kept the training and test set uniform across all these classifiers for each run.

E. Performance Evaluation

We compare the classification performance of each approach based on the AUROC since it allows for meaningful comparisons even in the presence of high class imbalance. As previously noted, feature engineering is of particular importance when the number of training examples is small. To this end, in our evaluation we varied the fraction of training data utilized, f, in each run described above. Training data were repeatedly randomly subsampled for each setting of f. Figure 5 shows the resulting average performance of each classifier on the test data, as we vary f. We plot the median values of AUROC along with error bars representing the inter-quartile ranges (IQR) (i.e., the first and the third quartile).

For each task, the HPAM+EMD model outperforms LDA and the other baselines when the number of training samples are limited (i.e., when $f \leq 0.1$). Also, we see that the HPAM+EMD approach consistently outperforms HPAM and LDA at all levels of training data used, underlining the significance of accounting for the distances between topics.

V. Discussion & Conclusion

In this paper, we address two issues that often arise when applying machine learning and data mining techniques to categorical data:

1) the high-dimensionality of the data results in sparse feature representations that make it challenging to identify meaningful similarities, and
2) the complexity of the underlying relationships among categories mean that standard Euclidean distance metrics often do not apply.

To address these issues, we propose a feature-learning framework that automatically incorporates expert knowledge. Our approach has two stages. In the first stage we learn a lower-dimensional feature representation of the data. Using hierarchical topic modeling applied to unstructured data (representing expert knowledge) we identify similar categories and collapse these categories into topics. These topics are the basis for the lower-dimensional feature space. The second stage accounts for the non-uniform distances between topics/features by incorporating the ground distance between topics (i.e., the distribution over the vocabulary) when calculating distance/similarity between examples in the lower-dimensional space.

We demonstrate the utility of our approach through a case study of patients and their medications. It is important to note that the goal of this case study was not to build the best possible classifier to predict the adverse patient outcomes, but to test the utility of the abstractions. The proposed feature-learning framework could be extended in a number of different

ways. For example, one could incorporate expert knowledge in the form of structured data. Also if the data were time-varying, temporal topic models could be used to discover topics that change across the period.

In the application we considered, the categorical data had a straightforward textual representation. However, our proposed feature-learning framework applies more generally (e.g., to the analysis of demographic features of different countries where the categories can be related to any trait of the study population). As with most dimensionality reduction techniques, we note that the learned abstractions help only in settings where the training data are limited. Despite the growing trend toward massive datasets, in many applications (particularly in healthcare) one still encounters a small number of examples. When the number of examples is small, but the dimensionality is high, the proposed feature-learning framework can help efficiently incorporate auxiliary expert knowledge.

ACKNOWLEDGEMENTS

This material is based upon work supported by the National Science Foundation Graduate Research Fellowship under Grant No. (NSF CN 1330142)

REFERENCES

[1] G. Hughes, "On the mean accuracy of statistical pattern recognizers," *IEEE Transactions on Information Theory*, vol. 14, no. 1, pp. 55–63, 1968.

[2] A. Jain, "Dimensionality and sample size considerations in pattern recognition practice," *Handbook of Statistics, Krishnaiah*, vol. 2, pp. 835–855, 1982.

[3] S. A. Caraballo, "Automatic construction of a hypernym-labeled noun hierarchy from text," in *Proc. of the 37th Annual Meeting of the Association for Computational Linguistics on Computational Linguistics*, 1999, pp. 120–126.

[4] S. P. Ponzetto and M. Strube, "Deriving a large scale taxonomy from Wikipedia," in *AAAI*, vol. 7, 2007, pp. 1440–1445.

[5] W. Dakka and P. G. Ipeirotis, "Automatic extraction of useful facet hierarchies from text databases," in *ICDE*, 2008, pp. 466–475.

[6] R. Navigli, P. Velardi, and S. Faralli, "A graph-based algorithm for inducing lexical taxonomies from scratch," in *IJCAI*, 2011, pp. 1872–1877.

[7] C. H. Ding, "A probabilistic model for dimensionality reduction in information retrieval and filtering," in *Proc. of 1st SIAM Computational Information Retrieval Workshop*, 2001.

[8] M. W. Berry, S. T. Dumais, and G. W. O'Brien, "Using linear algebra for intelligent information retrieval," *SIAM review*, vol. 37, no. 4, pp. 573–595, 1995.

[9] S. T. Roweis and L. K. Saul, "Nonlinear dimensionality reduction by locally linear embedding," *Science*, vol. 290, no. 5500, pp. 2323–2326, 2000.

[10] D. Mimno, W. Li, and A. McCallum, "Mixtures of hierarchical topics with Pachinko allocation," in *ICML*. ACM, 2007, pp. 633–640.

[11] Y. Rubner, C. Tomasi, and L. J. Guibas, "The earth mover's distance as a metric for image retrieval," *IJCV*, vol. 40, no. 2, pp. 99–121, 2000.

[12] Y. Rubner, C. Tomasi, and L. J. Guibas, "A metric for distributions with applications to image databases," in *ICCV*, 1998, pp. 59–66.

[13] P. Zhang, F. Wang, J. Hu, and R. Sorrentino, "Towards personalized medicine: Leveraging patient similarity and drug similarity analytics," *AMIA Joint Summits on Translational Science*, 2014.

[14] A. Uyar, A. Bener, H. N. Ciray, and M. Bahceci, "A frequency based encoding technique for transformation of categorical variables in mixed IVF dataset," in *EMBS, IEEE*, 2009, pp. 6214–6217.

[15] M. Tenenhaus and F. W. Young, "An analysis and synthesis of multiple correspondence analysis, optimal scaling, dual scaling, homogeneity analysis and other methods for quantifying categorical multivariate data," *Psychometrika*, vol. 50, no. 1, pp. 91–119, 1985.

[16] C. Faloutsos, *Searching multimedia databases by content*, 1996, vol. 3.

[17] M. Flickner, H. Sawhney, W. Niblack, J. Ashley, Q. Huang *et al.*, "Query by image and video content: the QBIC system," *Computer*, vol. 28, no. 9, pp. 23–32, 1995.

[18] Z. Huang, "Extensions to the k-means algorithm for clustering large data sets with categorical values," *Data Mining and Knowledge Discovery*, vol. 2, no. 3, pp. 283–304, 1998.

[19] A. Ahmad and L. Dey, "A k-mean clustering algorithm for mixed numeric and categorical data," *Data and Knowledge Engineering*, vol. 63, no. 2, pp. 503–527, 2007.

[20] J. Ji, W. Pang, C. Zhou, X. Han, and Z. Wang, "A fuzzy k-prototype clustering algorithm for mixed numeric and categorical data," *Knowledge-Based Systems*, vol. 30, pp. 129–135, 2012.

[21] B. Vrijens, S. De Geest, D. A. Hughes, K. Przemyslaw, J. Demonceau *et al.*, "A new taxonomy for describing and defining adherence to medications," *British Journal of Clinical Pharmacology*, vol. 73, no. 5, pp. 691–705, 2012.

[22] H. Saitwal, D. Qing, S. Jones, E. V. Bernstam, C. G. Chute *et al.*, "Cross-terminology mapping challenges: a demonstration using medication terminological systems," *Journal of Biomedical Informatics*, vol. 45, no. 4, pp. 613–625, 2012.

[23] V. Kashyap, "Design and creation of ontologies for environmental information retrieval," in *Proc. of the 12th Workshop on Knowledge Acquisition, Modeling and Management*, 1999, pp. 1–18.

[24] N. Pernelle, M.-C. Rousset, and V. Ventos, "Automatic construction and refinement of a class hierarchy over semi-structured data." in *Workshop on Ontology Learning*, 2001.

[25] R. Navigli, P. Velardi, A. Cucchiarelli, and F. Neri, "Extending and enriching wordnet with ontolearn."

[26] P. Velardi, S. Faralli, and R. Navigli, "Ontolearn reloaded: A graph-based algorithm for taxonomy induction," *Computational Linguistics*, vol. 39, no. 3, pp. 665–707, 2013.

[27] W. Wu, H. Li, H. Wang, and K. Q. Zhu, "Probase: A probabilistic taxonomy for text understanding," in *Proc. of the 2012 ACM SIGMOD International Conference on Management of Data*, 2012, pp. 481–492.

[28] H. Poon and P. Domingos, "Unsupervised ontology induction from text," in *Proc. of the 48th annual meeting of the Association for Computational Linguistics*, 2010, pp. 296–305.

[29] L. Dietz and A. Stewart, "Utilize probabilistic topic models to enrich knowledge bases," in *Proc. of the ESWC 2006 Workshop on Mastering the Gap: From Information Extraction to Semantic Representation*, 2006.

[30] W. Li and A. McCallum, "Pachinko allocation: DAG-structured mixture models of topic correlations," in *ICML*. ACM, 2006, pp. 577–584.

[31] G. B. Dantzig, "Application of the simplex method to a transportation problem," *Activity Analysis of Production and Allocation*, vol. 13, pp. 359–373, 1951.

[32] P. J. Moreno, P. P. Ho, and N. Vasconcelos, "A Kullback-Leibler divergence based kernel for svm classification in multimedia applications," in *Advances in Neural Information Processing Systems*, 2003.

[33] A. Zamolotskikh and P. Cunningham, "An assessment of alternative strategies for constructing emd-based kernel functions for use in an svm for image classification," in *IEEE International Workshop on Content-Based Multimedia Indexing*, 2007, pp. 11–17.

[34] S. Gu and Y. Guo, "Learning svm classifiers with indefinite kernels." in *AAAI*, 2012.

[35] V. Roth, J. Laub, M. Kawanabe, and J. M. Buhmann, "Optimal cluster preserving embedding of nonmetric proximity data," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 25, no. 12, pp. 1540–1551, 2003.

[36] M. Saeed, M. Villarroel, A. T. Reisner, G. Clifford, L.-W. Lehman *et al.*, "Multiparameter intelligent monitoring in intensive care II (MIMIC-II): a public-access intensive care unit database," *Critical Care Medicine*, vol. 39, no. 5, p. 952, 2011.

[37] G. Heinrich, "A generic approach to topic models," in *Machine Learning and Knowledge Discovery in Databases*, 2009, pp. 517–532.

[38] H. M. Wallach, "Topic modeling: beyond bag-of-words," in *ICML*. ACM, 2006, pp. 977–984.

[39] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei, "Hierarchical Dirichlet processes," *Journal of the American Statistical Association*, vol. 101, no. 476, 2006.

[40] T. Minka, "Estimating a Dirichlet distribution," 2000.