Journal of Data Science 15(2015), 323-340

Topic Model Kernel Classification With Probabilistically Reduced Features

Vu Nguyen*, Dinh Phung, and Svetha Venkatesh

Centre for Pattern Recognition and Data Analytics (PRaDA) Deakin University, Melbourne, Australia

Abstract: Probabilistic topic models have become a standard in modern machine learning to deal with a wide range of applications. Representing data by dimensional reduction of mixture proportion extracted from topic models is not only richer in semantics interpretation, but could also be informative for classification tasks. In this paper, we describe the Topic Model Kernel (TMK), a topicbased kernel for Support Vector Machine classification on data being processed by probabilistic topic models. The applicability of our proposed kernel is demonstrated in several classification tasks with real world datasets. TMK outperforms existing kernels on the distributional features and give comparative results on nonprobabilistic data types.

Key words: Topic Models, Bayesian Nonparametric, Support Vector Machine, Kernel Method, Classification, Dimensionality Reduction

1. Introduction

Data representation is critical in data analysis tasks. Central to Support Vector Machines are kernels, which maps the input data to another dimensional spaces in which the linear separating hyperplanes are easier to construct. Given a mapping function φ and two data points (x_i, x_j), the kernel function k computes inner product k φ (x_i), φ (x_j) without explicit computation of k(φ (x_i)) and k(φ (y_i)) separately. Several kernels have been introduced in literature that has examined appropriate kernels for a wide variety of data. Each dataset requires careful choice of the appropriate kernel for SVM classification. In this paper we focus on a class of problem for SVM when the feature input can be conveniently represented in distributional forms. Such distributions constitute rich information one can exploit, as they are outputs from the probabilistic topic models Blei et al. (2003) whose latent variables can be used as distributional representation for data. Examples include Probabilistic Latent Semantic Indexing (PLSI) Hofmann (1999), Latent Dirichlet Allocation (LDA) Blei et al. (2003) or Hierarchical Dirichlet Processes (HDP) Teh et al. (2006), which can produce multinomial distributions over topics given text data or raw pixels in images. This representation is not only richer in semantics than the original bag of words, but also Blei et al. (2003) have demonstrated that the topic model features could be more informative for classification than the raw word feature as demonstrated in Blei et al. (2003). Moreover, such derived features occupy only 0.04 percent in space compared to a very large raw feature set of individual words.

The combinations of generative approaches (such as LDA, HDP) with discriminative ones (e.g. SVM) have recently shown to be very effective Fritz and Schiele (2008); Phung et al. (2012). Hence it is attractive to expose methods integrating these statistical models and discriminative classifiers. Furthermore, we are motivated by recent successful applications of Jensen Shannon divergences to compute the similarities and distances when the data are drawn from probabilistic distributions Antolín et al. (2009); Wartena and Brussee (2008); Nguyen et al. (2013a).

In this paper, firstly we make use of preprocessing raw data by topic models for extracting the latent feature in probabilistic space. The probabilistic feature is then utilized for classification task. We propose of a proper kernel originated from the Jensen-Shannon divergence Endres and Schindelin (2003), namely Topic Model Kernel (TMK)[1] for optimizing the discriminative among these features. The source code is released at the first author webpage[2]. The recent advance in Bayesian nonparametric modelling, such as the HDP Teh et al. (2006) which automatically determine the number of topics, make the proposed classification framework more attractive to real-world application. We conducted extensive experimental validation of the proposed TMK which outperforms other existing kernels on the probabilistic derived features and yields a comparative performance on *other data types* (non-distribution guarantee).

2. Related Work

2.1 Loss reserves data for Israel

Support Vector Machines (SVM) Cortes and Vapnik (1995) is a very well-known supervised learning method for classification. The SVM optimization equation Boser et al. (1992); Chang and Lin (2011) for binary case is expressed as:

$$\min_{w,b,\xi} \frac{1}{2} W^T + C\sum_{i=1}^{l} \xi_i \tag{1}$$

$$\text{Subject to } y_i(W^T\phi(x_i + b) \geq 1 - \xi_i$$

$$\xi_i \geq 0$$

where (x_i,y_i) is a set of instance-label pair; $x_i \in R^n$ and $y \in \{1, -1\}^l$ and ξ_i is a slack variable. For muticlass SVM, it aims to assign labels y ∈ 1, 2, 3, ... *m* to each instance which is typically reduced the single multiclass issue into multiple binary classification tasks. A mapping function ϕ *(x)* here becomes $X \rightarrow R^M$.

SVM is laying within a broader umbrella of kernel methods Shawe-Taylor and Cristianini (2004) that approaches the supervised learning problem by mapping the data into a high dimensional feature space. The goal is to find a better representation by this mapping transformation. Because the mapping can be general, there are numerous existing kernels in literature, including Exponential Kernel, Laplacian Kernel, Inverse Multiquadric Kernel, Cauchy Kernel, and so on. Each kernel is taking into account for different 'genres' of the real world data. Some examples of kernel functions are summarized below.

- Radial Basic Functyion Kernel (RBF) or Gaussian Kernel:

  $$k(x, y) = exp(-\gamma\|x - y\|^2).$$

  It is recommended as the first choice for Support Vector Machine (Chang and Lin (2011)). The parameter γ plays a crucial role in the classification performance.
- Linear Kernel : $K(x, y) = x^T y + c$ where c is a constant.
- Ploynomial Kernel: $k(x, y) = (\alpha x^T y + c)^d$ with polynomial degree d.
- Sigmoid Kernel: $k(x, y) = tanh(\alpha x^T y + c)$ where slope parameter $\boldsymbol{\alpha}$ needs to be adjusted for the best performance.
- Inverser Multiquadric Kernel: $\mathrm{k}(\mathrm{x}, \mathrm{y}) = \frac{1}{\sqrt{\|x-y\|^2 + C}}$ where $\boldsymbol{c}$ is a constant.
- Power Kernel: The Power kernel is also known as the (unrectified) triangular kernel. It is an example of scale-invariant kernel Fleuret and Sahbi (2003) and is also only conditionally positive definite Boughorbel et al. (2005): $k(x, y) = -\|\mathrm{x} - \mathrm{y}\|^{-\beta}$ where $\boldsymbol{\beta}$ is a parameter from $0 < \boldsymbol{\beta} < 1$
- Spline Kernel: the Spline kernel is given as a piece-wise cubic polynomial, as derived in the works by Gunn (1998). With $x,y \in R^d$, we have:

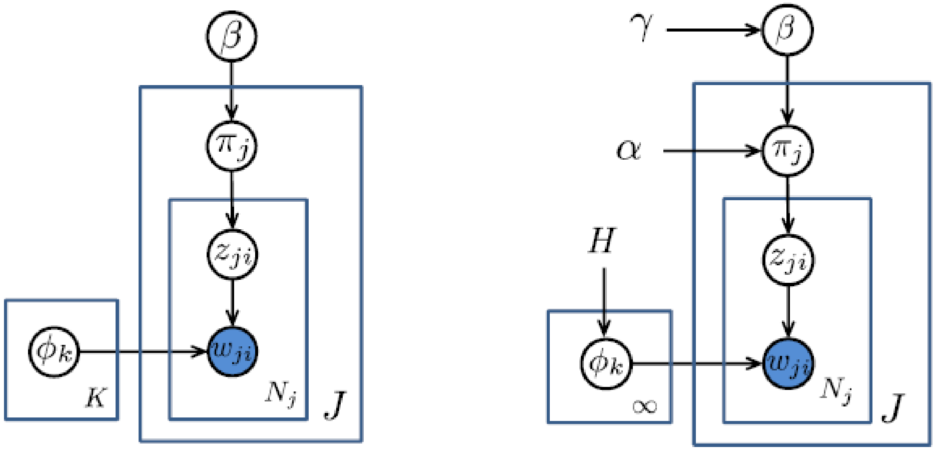
$$k(x,y) \; = \prod_{i=1}^{d}(1 + x_i y_i + x_i y_i min(x_i, y_i) + \frac{x_i + y_i}{2} min(x_i + y_i)^2 + \frac{min(xi,yi)^3}{3}).$$

- Cauchy Kernel: the Cauchy kernel comes from the Cauchy distribution Basak (2008). It is a long0tailed kernel and can be used to give long-range influence and sensitivity over the high dimension space. The kernel is defined by the kernel function with smoothing parameter **σ**.

$$\mathrm{k(x, y)} = \frac{1}{1 + \frac{\|x - y\|^2}{\sigma^2}}$$

Kernel selection is heavily dependent on the data types. For instance, the linear kernel is important in large sparse data vectors and it can be seen as the simplest of all kernels. Whereas, the Gaussian (or RBF) are general purpose kernels used when prior knowledge about data is not available. It decreases with distance and ranges between 0 (in the limit) and 1 (when x = y). The polynomial kernel is widely applied in natural language processing Goldberg and Elhadad (2008) while Spline Kernel is usually reserved for continuous-space image processing Horbelt et al. (2000). Because classification accuracy heavily depends on kernel selection, researchers had proposed to have kernel functions based on a general purpose learning and domain specific. A specific data type requires a suitable kernel for their best performance as working with SVM classification. The most appropriate kernel must guarantee the smoothness amongst data within the same class, maintain distinction to others classes. In this paper, we propose TMK for the probabilistic feature derived by topic models.

We are motivated by the importance of the low dimensional features derived by topic models. In real-world applications, e.g. text analysis, the raw data always are



(a) LDA Representation

(b) HDP Representation

Figure 1 : Probabilistic Topic Models.

There are *J* documents, each document has N_j (observed) words w_{ji}. Each word is assigned to a topic $\boldsymbol{\phi_k}$ (assumed there are *K* topics) through latent variables $Z_{ji} = k$. The mixture weight π_j - captures the distribution of $\{Z_j\}_{i=1}^{N_j}$ over each topic (1, … , K). $\boldsymbol{\beta}$ is a parameter for Dirichlet distribution to generate $\boldsymbol{\pi_j}$. $\boldsymbol{\gamma}$ and *H* serve as a hyper parameters for $\boldsymbol{\beta}$ and $\boldsymbol{\phi_k}$ (in HDP only), respectively.

represented in high dimensional, which the dictionary size can be thousand or hundred thousand dimension. Therefore, extracting the low dimensional hidden feature embedded inside the raw data is essential for richer in semantic and informative for classification.

We choose the four baseline kernels: RBF, Linear, Polynomial, and Sigmoid, for comparison with the proposed kernel. The four kernels, which are built-in in LibSVM Chang and Lin (2011), are being used extensively as a common choice for classification with SVM.

2.2 Probabilistic Topic Models

The discrete distribution features in practice can be the outcome from probabilistic topic models that has become popular in modern machine learning. At the first glance, the probabilistic mixture models, can be seen as mixture distribution, comprise an underlying set of distributions transforming the complex data into a group of simpler densities. Blei et al. (2003) introduce the topic model, Latent Dirichlet Allocation which is a class of topic models providing a simple way to analyze large volumes of unlabeled text. A 'topic' consists of the cluster of words that frequently occur together. There are K topics $\boldsymbol{\phi_k}$, k ∈ {1, ..., K} which are discrete distributions over words. For example, a topic 'sport' may contain high probabilities to such words as 'athlete', 'tennis', 'championship'. Then, each document is assumed to be characterized by a mixture of topics. Our focus is on document feature representation, the mixture proportion (the latent variable π_j on Figure 1) which is a k-dimensional vector. Each element k-th of vector π_j indicates how much the document *j* contributes to the topic k-th. Traditionally, we need to input the number of topic K for the model. However, Bayesian nonparametric models, such as Hierarchical Dirichlet Process Teh et al. (2006), can identify the suitable number of K. The good model guarantees to return the posterior distribution of the underlying expressive factors for the observed data.

These topic models (e.g., LDA, HDP) are designed to work with a single data channel (e.g. word observation in a document). To accommodate the additional context information (e.g., timestamp, location) Nguyen et al. (2014a) have recently proposed the Multilevel Clustering with Context model (MC2). To demonstrate our Topic Model Kernel, we consider the extracted feature from all of three settings: (1) traditional single observation in parametric (fixed number of topic), (2) single observation in nonparametric (the number of cluster is automatically identified), and (3) multiple observations in nonparametric setting (e.g. word, timestamp, location, etc). For single observation, there are noticeably Latent Dirichlet Allocation Blei et al. (2003) in parametric setting and Hierarchical Dirichlet Processes Teh et al. (2006) in nonparametric configuration. For multiple observation, we consider the Multilevel Clustering with Context (MC2) Nguyen et al. (2014a). The detailed generative processes and posterior inferences behind these prototypes can be found in the original papers Blei et al. (2003); Teh et al. (2006); Nguyen et al. (2014a). Essentially, the algorithm proceeds by looping iteratively through each of the data points and performing MCMC moves on the cluster indicators for each point.

2.2.1 Probabilistic Topic Models

LDA is a parametric model that can be described using a pre-defined and finite number of paramteres. The graphical representation of LDA is displayed in Figure 1a. In LDA model, there are *K* topics $\boldsymbol{\phi_k}$, $k \in \{1, \ldots, k\}$ (K is initialized and fixed), which are discrete distributions over words.

To extract the mixture components for a corpus of documents, posts, or images in a large scale dataset, the most common probabilistic topic models of Latent Dirichlet Allocation Blei et al. (2003) is taken into account. Particularly, the latent factor ϕ_k in Figure 1, is a characterized by a distribution over words and the mixture proportion outputs π_j in Figure 1a is a random mixtures over hidden topics that reflect the topic assignment distribution over each document. We describe the generative process of LDA below.

$$
\begin{aligned}
\phi_k &\sim \text{Dir}(\eta)\ \ k\in\{1,\ldots,K\}\\
\pi_j &\sim \text{Dir}(\beta)\ \ j\in\{1,\ldots,J\}\\
z_{ji} &\sim \text{Multinomial}(\pi_j)\ \ j\in\{1,\ldots,J\},i\in\{1,\ldots,N_j\}\\
w_{ji} &\sim \text{Multinomial}(\phi_{z_{ji}})\ \ j\in\{1,\ldots,J\},i\in\{1,\ldots,N_j\}
\end{aligned}
$$

According to the conjugate property of Multinomial-Dirichlet (Schlaifer and Raiffa (1961)), the posterior estimation of π j are assumed to be drawn from multiple group-specific distributions and this allows documents within a class share the same set of weights – this nature will be benefit in classification.

2.2.2 Hierarchical Dirichlet Processes

For demanding tasks such as computer vision, text modelling and understanding, it is not well-applicable to merely rely on parametric models due to their incompetence of capturing enough complexity of the vast continuing data. Hierarchical Dirichlet processes Teh et al. (2006) (See Figure 1b), an attractive Bayesian non-parametric framework with the crucial advantage to hierarchy modelling and naturally address the problem of model selection. HDP emphasizes in exploiting the sharing statistical property across documents. Alternatively from LDA, the mixture proportion π_j in HDP is generated from Dirichlet Process Ferguson (1973) with concentration parameter α and the base measure defined by β.

2.2.3 Multilevel clustering with context

In many situations, the data (e.g. image features, events) naturally is associated with additional context information (e.g. images annotations, time and location of events.) The context information is important because it offers the multi-view information toward the data. There, discovering the hidden structure embedded inside the data and context can be improve the classification performance rather

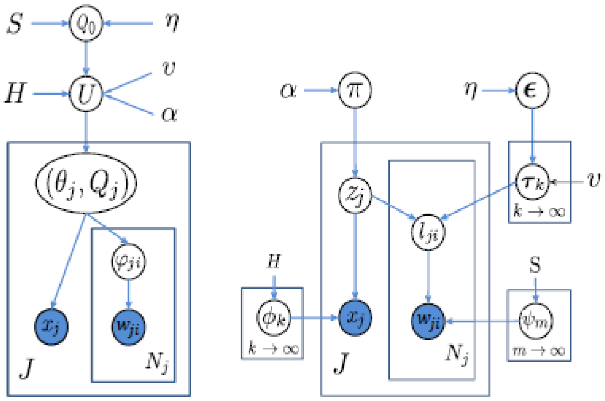


Figure 2: Graphical representation of MC2.

Each document has a context observation x_j and the content words w_{ji}(s). Hidden variable Z_j assigns a document j to a cluster k and context observation to a context topic ϕ_k and l_{ji} assigns content word to a content topic ψ m. τ k and ϵ contains the mixture proportion of assigning topic m in cluster k and overall cluster, respectively. α , η , v define the number of cluster and topics.

Than just focusing on the main data. Recently, Nguyen et al. (2014a) has proposed the Multilevel Clustering with Context (MC^2), a Bayesian nonparametric model to jointly cluster both context and groups while fully utilizing group-level context. The model flexibly adapts the data with and without context information. Using the Dirichlet Process as the building block, MC^2 constructs a product base-measure to accommodate both context and content data observations. We present the graphical representation of MC^2 in Figure 2. Details of model properties and inference refer to the original paper Nguyen et al. (2014a,b).

3. Topic Model Kernel

3.1 Kullback-Leibler Divergence

The Kullback–Leibler (KL) divergence Kullback and Leibler (1951), introduced in information theory and probability theory is a non-symmetric measure of the similarity between two probability distributions. Its intuitive understanding arises from likelihood theory Shlens (2007) measuring the distance between the initialized

$$D_{KL}(P \parallel Q) = \sum_i P(i) \ln \frac{P(i)}{Q(i)}$$

and for continuous distributions as:

$$D_{KL}(P \parallel Q) = \int_{-\infty}^{+\infty} p(x) \ln \frac{p(x)}{q(x)} dx$$

where *p* and *q* denote the densities of the distributions *P* and *Q*. Moreno et al. (2003); Chan et al. (2004) have proposed a symmetric KL divergence kernel for classifying objects under the generative model of Gaussian mixture, a step toward classifying distribution data with SVM.

3.2 Jensen-Shannon Divergence

Based on the KL divergence, the Jensen-Shannon (JS) divergence Endres and Schindelin (2003) calculates the distance between two probability distributions P and Q as:

$$D_{JS}(P,Q) = \pi D_{KL}(P \parallel M) + (1-\pi) D_{KL}(Q \parallel M)$$

where M = $\frac{1}{2}$ (P + Q) and D_{KL} is the KL divergence discussed in Section 3.1. The lower bound of JS divergence is 0 when the two distributions are identical. Its square root Endres and Schindelin (2003) is proof as an asymptotic approximation to the well-known χ^2 and being a metric with the triangle inequality property for two distributions. This distance can be seen (in the symmetric KL flavour) as the average distance between two random distributions to their empirical mean, with π is set as 0.5 Chan et al. (2004). Another interesting property of JS divergence is negative definite on R+ × R+ Topsoe (2003) that will be useful when we verify for kernel validation.

3.3 Topic Model Kernel

The kernel function is basically a measurement criteria that compares the similarity between two points or vectors. But not all of the measurement distances or similarity functions yield proper attributes to be a valid kernel. The Topic Model Kernel (TMK) is defined following:

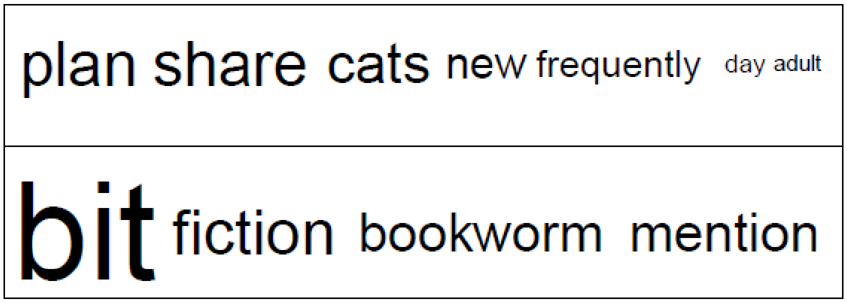

plan share cats new frequently day adult
bit fiction bookworm mention

Figure 3: Two examples of LDA topic ϕ_k on LiveJournal data.

where p and q denote the densities of the distributions P and Q. Moreno et al. (2003); Chan et al. (2004) have proposed a symmetric KL divergence kernel for classifying objects under the generative model of Gaussian mixture, a step toward classifying distribution data with SVM.

$$K_{TMK}(X,Y) = \exp\left\{-\frac{1}{\sigma^2}\times D_{JS}(X,Y)\right\}$$

$$= \exp\left\{-\frac{1}{\sigma^2}\times\left[\frac{1}{2}\sum_i X(i)\ln\frac{X(i)}{M(i)}+\frac{1}{2}\sum_i Y(i)\ln\frac{Y(i)}{M(i)}\right]\right\} \qquad (2)$$

By exponentating the negative JS divergence, it leads to the positive definite kernel function K_{TM} because (1) JS divergence is negative definite on R+ × R+ Topsoe (2003), (2) let exponentiate the negative of JS divergence giving the positive definite kernel that projecting the divergence distance into the bounded range of 0 and 1. Thus, TMK satisfies theMercer condition of $c^T K_{TMC} \geq 0$ with $K_{TM(i,j)} = K_{TM}it(x_i,x_j)$ for the validity of te kernel. The variance σ^2 plays a role as shape parameter to flexibly flat or widen the data.

4. Experiment Results and Analysis

Experiments are conducted using world data in various classification scenarios, including:

- The topic model features derived from single observation in parametric from of LDA or nonparametric counterpart as HDP.
- The extracted feature from multiples observation of MC2 model.
- The generic features are obtained from other sources that we do not guarantee them fit into any type of distribution
- We analyze the kernel performance on parameter space to verify our kernel’s superiority on probabilistic features.
- We demonstrate a possible way of classification as combined product of raw feature and topic model feature for better performance in classification.

We use the LibSVM Chang and Lin (2011) as a standard library to compare the proposed kernel with four baseline LibSVM built-in kernels: Radial Basic Function Kernel, Linear Kernel, Polynomial Kernel, and Sigmoid Kernel. The data will be scaled as recommended in LibSVM to ensure the best performance. We focus on the multi-class classification problem viewed as multiple binary classification problems.

The scores are reported at two types of parameter: the default parameter (set by LibSVM) and the optimal parameter by brute-force cross validation searching (as the default parameter sometimes cannot provide the best performance). For Topic Model Kernel, we empirically set the default parameter σ^2 is equal to the feature dimension size after observing TMK operations on several datasets. Throughout this experiment, the whole data is randomly splited into 10 sets, which comprise of training set and testing set such that the instances in testing is not appearing in training set.

4.1 Topic Model Features

LDA and HDP are used to model the single observation data (e.g. words in a document). We run LDA and HDP to extract the mixture proportions π j on Livejournal, Reuter21578, and LabelMe dataset. LDA is carried out on Live Journal and Reuter21578 datasets and HDP on LabelMe to extract the mixing proportion features, then use SVM for classification with the proposed kernel.

4.1.1 Livejournal DataSet

Data processing set up: We crawled the communities listed in the Livejournal directory, retrieved August 2012. These communities are categorised by Livejournal into 10 categories from the 100 communities obtained, summarizing of 8,758 posts giving the vocabulary size of 65,483 which is the feature dimension of raw data. The task is to predict the category, given text data from user's posts. We treat each user post as a document and run LDA with fixed number of latent factors from {6, 10, 20, 50}. Latent Dirichlet Allocation is carried for the whole dataset with 1000 iteration Gibbs sampling. The examples of estimated topic φ_k , about literature and life, are visualized in Figure 3 and our LDA features are in Figure 4 which reduced from original high dimension of 65,483 to 50.

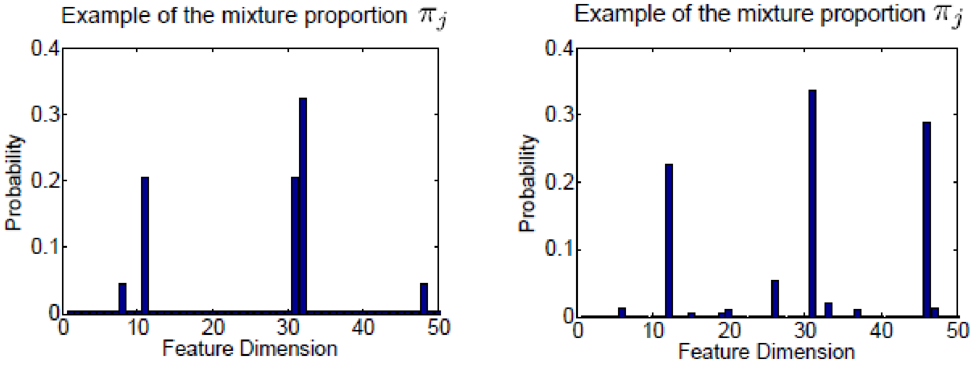


Figure 4: Two examples of the reduced feature π_j by LDA from 65,483 to 50.

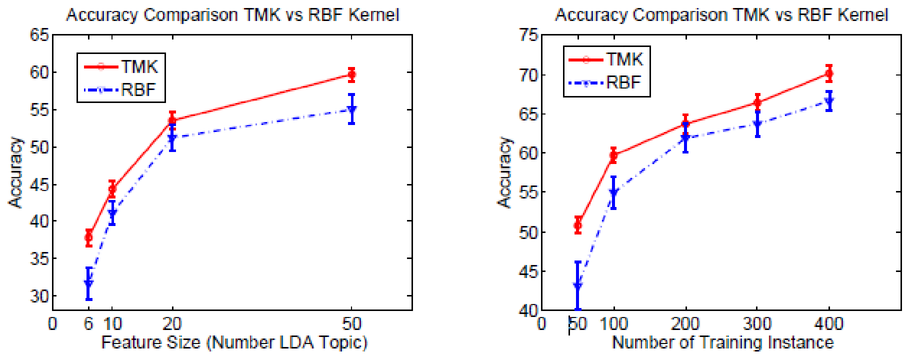
Classification set up: We do the experiments progressively with increasing numbers of training instances from 10 to 400 (refer Figure 5b) and varying the number of hidden factors K (refer Figure 5a). The optimal parameter (for the best performance) is achieved with 3 fold validation on training data sets. The performance is judged by averaging 10 random subsets of train/ test datasets.

The results in Figure 5 and Table 1 demonstrate the superiority of our kernel and clearly shows the effect of increasing the number of learned feature or number of training instances.

4.1.2 Reuter21578 Datase

Data prcessing set up: Reuter21578 is a common dataset for text classification. It consists of documents appeared on the Reuters Newswire in 1987. There are totally 10 categories for classification. Similar to Live Journal data, we utilize posterior inference of LDA on Reuters21578 dataset (again using 1000 iterations of Gibbs sampling) to extract the mixing proportion feature π_j in which the number of hidden factors is set as K = 20.

Classification set up: The accuracy comparison is displayed in Table 1. The Topic Model Kernel (TMK) outperforms four baseline kernels on this dataset in both cases of parameter (default and optimal). The number of training instances and testing instances are set as 100 instance for each category (totally 1000 instances for training and 100 instances for testing). The final classification score is reported with standard deviation in 10 randomly experiment subsets. We observed that the optimal parameter for SVM along with kernel feature, obtained by brute-force searching, slightly increases the accuracy about 1% on LDA feature, whereas in other types of data, the input parameters will make a significant effect on the accuracy Hsu et al. (2003)



(a) With varied feature dimension. (b) With varied number of training data.

Figure 5: experiments comparison between TMK and RBF kernels on LDA feature derived from Live Journal Data

Table 1: Accuracy comparison of SVM classification on features derived from LDA.

Accuracy	Default Parameter		Optimal Parameter	
Kernel	LiveJournal	Reuter21578	LiveJournal	Reuter21578
TMK	**58.10**±2.15	**81.33**±0.20	**58.70**±1.78	**81.87**±0.17
RBF	54.90±4.93	79±0.55	55.00±4.85	79.40±0.55
Linear	54.40±5.29	78.27±0.13	54.90±4.28	79.07±0.13
Polynomial	52.60±6.65	77.93±0.10	54.20±5.20	78.93 + 0.10
Sigmoid	51.80±5.18	77.40±0.48	53.50±4.79	79.20±0.34

4.1.3 LableMe Dataset

Data processing set up: LabelMe Oliva and Torralba (2001) is the well-known benchmark dataset for image annotations and object categorizations that contents a bunch of images and tags. The subset of 8 classes LabelMe is justified for this experiment including ground truth 8 categories classification consisting of tall buildings, inside city, street, highway, coast, open country, mountain, and forest in totally 2688 images. To discard the noise and mistagging issues, top 30 high frequency tags are chosen giving a vocabulary size of 30. The Hierarchical Dirichlet Processes Teh et al. (2006) is carried out to extract the topic assignment feature flexibly, each image is treated as a document while each tag is considered as a word w_{ji} (refer Figure 1b) in the model. The collapsed Gibbs inference during 500 iterations are collected to compute the posterior. HDP automatically identifies 24 topics $\boldsymbol{\phi_k}$, four of whom is displayed in Figure 6 for visualization. The extracted feature π_j by HDP is therefore under the dimension of 24 (see Figure 7 for two examples) where two different classes are likely to gave dissimilar feature $\boldsymbol{\pi_j}$

Classification set up: The evaluation procedure is conducted alike the previous experiments that we splits the data into 10 trainning and testing subsets. In each subset, there are 800 and 800 instances for training and testing erespectively (100 instance in each calss). Then we run 3 fold cross validation to get the optimal parameter for testing. The performances with default SVM parameter (the default σ^2. Due to the sparsity of image tag and extracted feature, Linear kernel achieves the best performance at the default parameter. However, our TMK attains the best performance at the optimal parameter.

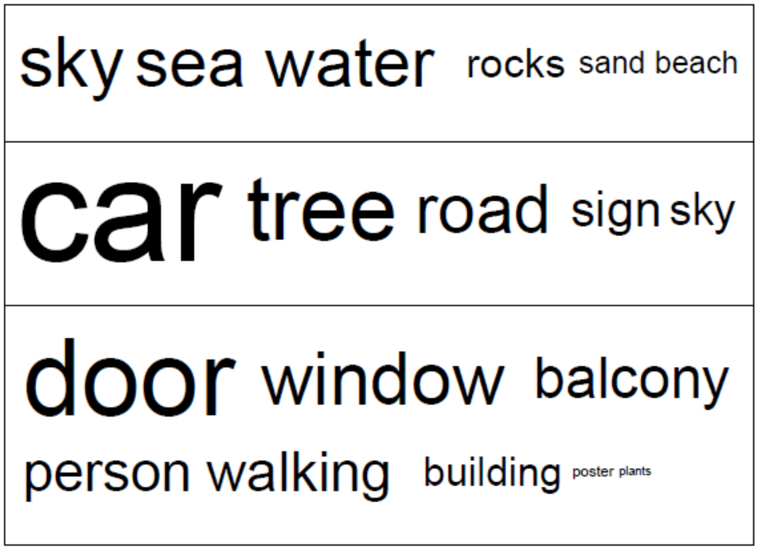
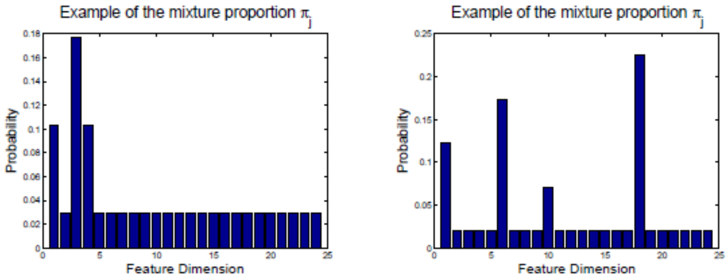

sky sea water rocks sand beach
car tree road sign sky
door window balcony
person walking building poster plants

Figure 6: LabelMe dataset: the learned topics ϕ_k by HDP.



a) Highway class. b) Buildings class.

Figure 7: Two examples π_j of the HDP feature on LabelMe dataset.

Table 2: Classification comparison on LabelMe dataset from features learned by HDP.

Accuracy	TMK	RBF	Linear	Polynomial	Sigmoid
Default Parameter	72.3±1.96	73.5±1.99	**74.5±1.87**	62.7±4.26	72.8±1.66
Optimal Parameter	**76.1**±1.88	73.3±2.08	73.8±1.46	74.5±2.22	73.9±1.30

4.2 Topic Model Features From Multiple Observations Model

Priviously, we have illustrated experiments on the feature derived from LDA and HDP running on single observation dataset. In this section, we aim to learn the performance of the topic model feature extracted from model with utilizing information from multiple observation (e.g. word, timestamp, authors in a document) for comparison.

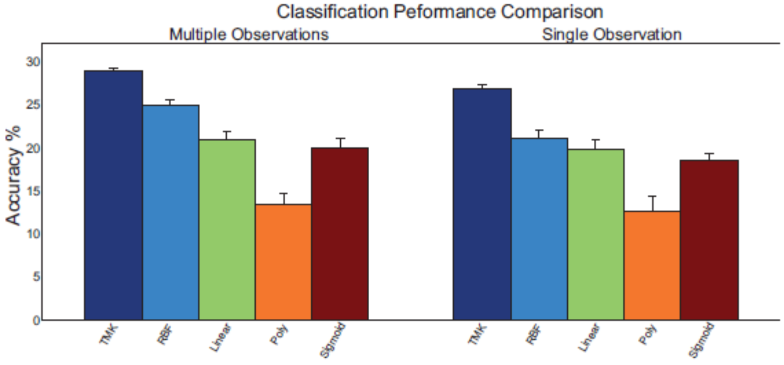


Figure 8: Classification accuracy comparison on NUS-WIDE dataset from featured learned by the Miltilevel Clustering with context in two settings: multiple observation (with context information) and single observation (without context information).

NUS-WIDE DATASET

Data processing set up: We set up to run Multilevel Clustering with Context (MC^2) Nguyen et al. (2014a) on NUS-WIDE subset of the 13-class animals with 2054 images. We use the available image label for image classification task. The feature vector includes 1000-dim annotation and 500-dim bag-of-word SIFT. We consider two cases of experiment on MC2 model : (1) running multiple observations of both annotation and SIFT (multiple observation), and (2) running the model on single observation of SIFT only (single observation). The posterior inference of MC^2 returns 15 topics. The topic model feature for each image in this MC^2 is not directly obtain like LDA or HDP. We compute the mixture proportion feature for each image using the latent indicator (variable l_{ji} in Figure 2). The mixture proportion π_j for an image j-th is a vector in M-dimension, where M-dimension, where M is th e number of topic discovered by MC^2. Each elements π_{jm} is computed in such a way similar to HDP Teh etal. (2006).

Classification set up: The classification comparison between multiple kernels is displayed in Figure 8 where we use 100 images per class (totally 1300 images) for training and 50 images per class (totally 650 images) for testing. Standard deviation error is calculated across 10 randomly experiments. Our kernel achieves the best performance among other kernels. Figure 8 presents the scores at default parameter (optimal parameter yields similar performance). The classification performance from the feature obtained by multiple observation slightly better than single observation. The extracted features from multiple data-source model can be richer in semantics and more informative for classification than the feature from single data-source model counterpart. There are two proper reasons for this claim. The first reason is that the context information from multiple data source prevents the model from over-fitting to the single data source. Another reason is the additional data channel offering the multi-view information toward the data instances in different categories. Therefore, jointly modelling multiple data observation produces informative features which are improving performance for classification.

Figure 9: Examples of digit 3 and 0 in MNIST dataset.

4.3 Non-distributional Data Source

To highlight the applicability of the TMK, we show how the proposed kernel performs on the raw data of MNIST dataset instead of extracting topic model features as previously. This experiment is aiming to discover the wide applicability of TMK on such kind of non-distribution data.

MNIST DATASET

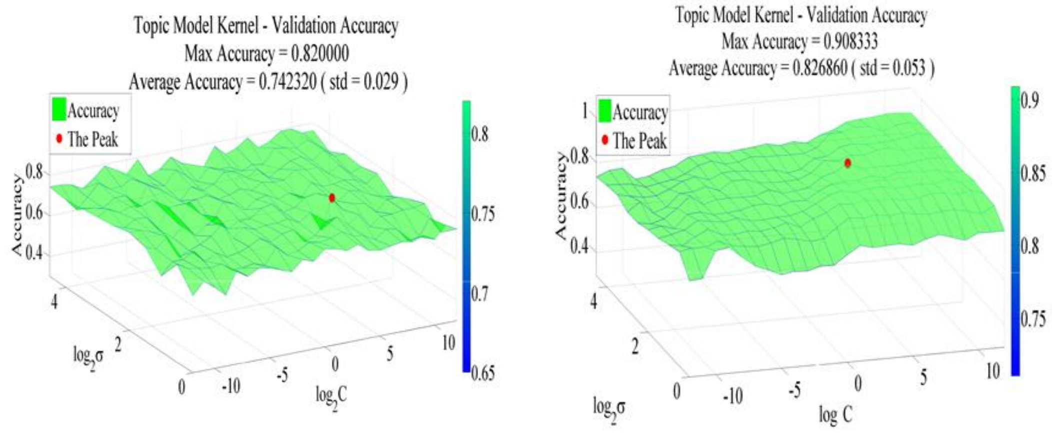
Data processing set up: The MNIST dataset LeCun et al. (1998) is a well-known dataset of handwritten digits, referred as a standard benchmark for many tasks, especially in classification problem. The ready-to-use extracted feature is available at author website (http://yann.lecun.com/exdb/mnist/) with the classification performance and the state of the art result on 60,000 training and 10,000 testing instances. In this experiment setting, we do not aim to beat the state of the art result on MNIST, but we want to illustrate the classification comparison between the TMK versus others with SVM tool.

Classification set up: We randomly pick up 100 items for training and another 100 for testing set and run for 10 times. We do not run for the whole 60,000 training vs 10,000 testing due to (1) resource limitation when constructing the gram matrix of the huge data (2) our goal is to proof the efficiency of the TMK by comparing with other kernels, not strike the state of the art result. The feature dimension of each gray image is 784 (28×28 pixels) at which the pixel value ranges from 0 to 255 (refer Figure 9 for examples). We note that this kind of raw image data is not pledged to drawn from any type distribution when use with Topic Model Kernel for classification. The accuracy is displayed in Table 3, although Linear kernel perform very well with default parameter, our kernel achieves the best result in optimal parameter (with brute force searching on validation set). The detailed performance on parameter space of MNIST dataset is discussed in the next section.

4.4 Parameter Selection Analysis

We now move on to our characterization of performance on various axes of parameters. To demonstrate the TMK is more robust on the parameter space, we record the accuracy planes with parameter of C in equation 1 for SVM and TMK parameter σ shown in Figure 10a). We get the peak accuracy of 0.82 on by 3 fold cross validation at which the optimal parameter is further used for testing. The average accuracy with standard deviation is used to evaluate the preeminent of TMK when the data is drawn from distribution. Topic Model Kernel accomplishes the best in the way that it get the highest score on average accuracy (0.74), lowest standard deviation (0.029), and the TMK's peak (0.82) is the highest among four baseline kernel's peaks (refer Table 4). Detailed visualization performances of the baseline kernels on HDP feature are illustrated in Figure 11. We observe that RBF, Linear, and Sigmoid kernels are quite settled than Polynomial kernel.

Further, we would like to see the performance on the non-distribution feature when varying the parameters of TMK. Although it is not really stable (with high standard deviation and lower average accuracy on the grid), it performs pretty well with comparable accuracy to other kernels at a certain area (can be obtained by cross validation).



(a) HDP feature on LabelMe dataset.↵ (b) Raw feature on MNIST dataset.↵

Figure 10: Topic Model Kernel cross validation accuracy by brute-force parameter searching.

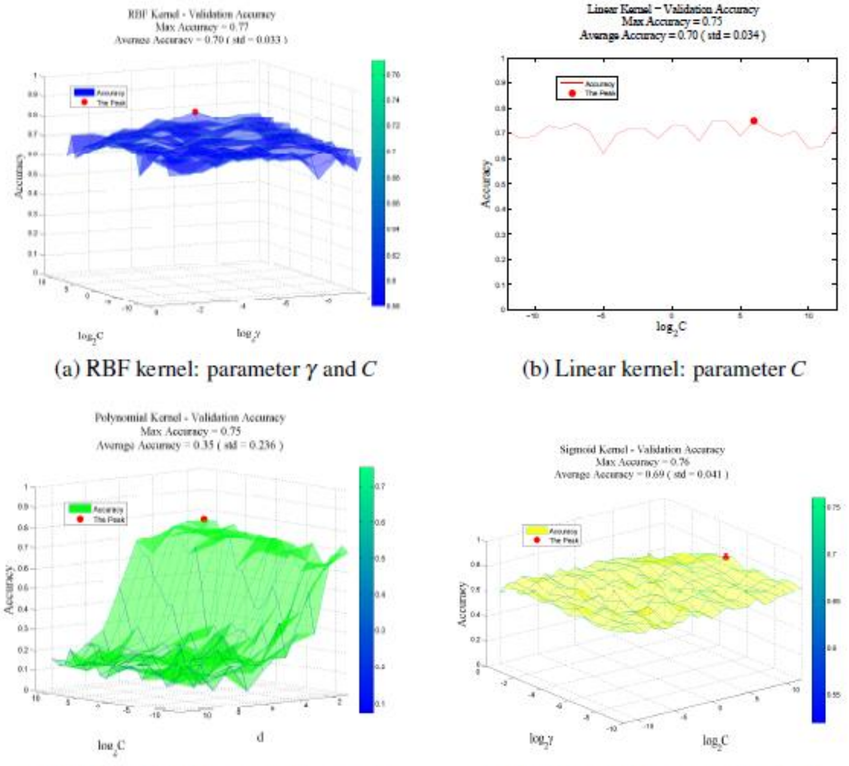
Table 4: Cross validation accuracy on parameter space comparisons of probabilistic feature of HDp versus non-probabilistic feature (or raw feature).

Kernel	LabelMe: HDP Feature			MNIST: Raw Feature		
	Peak	Average	Std	Peak	Average	Std
TMK	**0.82**	**0.74**	**0.029**	**0.91**	**0.83**	0.053
RBF	0.77	0.70	0.033	0.90	0.67	0.252
Linear	0.75	0.70	0.034	0.88	0.80	**0.029**
Polynomial	0.75	0.35	0.236	0.88	0.44	0.295
Sigmoid	0.76	0.69	0.041	0.85	0.43	0.304

4.5 Improved Classifiction Performance with Feature Combination

To analyze the classification performance under different feature kinds, we compare performances with various features including raw feature, extracted feature by HDP, extracted feature by MC2. Here, we use the MC2 with single observation (without context information) to be fair classification comparison with HDP and raw feature. In addition, we want to improve the classification by using feature produced by combining these individual features. The mixture proportion extracted (from raw data) by topic models offers an additional view to the data. It captures information from statistical perspective, representing proportion over underlying topics Blei et al. (2003). By focusing on the underlying topics, the topic model features ignore the noise information from the data. Two documents in the same class would likely to have similar mixture proportions.

We perform experiments on NUSWIDE dataset (similar to Section 4.2) with various features for comparison (refer Figure 5). HDP and MC2 models produce features which get similar performance for classification. MC2 feature slightly excesses HDP feature in classification but it is not distinction. We use MC2 model with annotation as the additional context information (multiple observations case). The raw feature itself attains better classification performance compared to features extracted by HDP and MC2. The possible reason for it is that the raw feature dimension is 500 while HDP and MC2 features are only 15. The higher dimension feature contains richer information toward the data. Furthermore, the features combined by Raw+HDP and Raw+MC2 achieve the best performance for classification. The joined features are better than the raw feature itself and the topic model feature (HDP or MC2) individually. Here, we do not include experiment from LDA because the performance of LDA can be seen from HDP (at the fixed number of topic). HDP is a Bayesian nonparametric counterpart of LDA, e.g. HDP Teh et al. (2006) automatically identifies the suitable number of topics while LDA does not. If we fix the number of topics, it is well-known from topic model literature that HDP and LDA would yield similar features. Our proposed kernel demonstrates its superior performance on these features comparing to other baseline kernels. In our experiment, we found that the combined feature between the raw feature and the topic model feature produce the best classification performance.



(a) RBF kernel: parameter γ and C

(b) Linear kernel: parameter C

(c) Polynomial kernel: parameter d and C

(d) Sigmoid kernel: parameter γ and C

Figure 11: Accuracy on parameter space of four baseline kernels on HDP feature of LabelMe dataset.

Table 5: Classification with different features on NUSWIDE dataset.

Feature	RBF	Linear	Poly	Sigmoid	TMK
Raw	**32.98±0.3**	30.82±0.2	9.72±0.6	32.85±0.3	32.75±0.2
HDP	29.29±0.7	27.95±0.4	19.17±1.3	28.48±0.5	**30.20±0.5**
MC2	29.31±1.1	27.57±1.5	18.22±1.0	27.09±0.3	**31.95±0.3**
Raw+HDP	33.09±0.4	31.31±0.3	10.12±1.2	32.86±0.4	**37.73±0.3**
Raw+MC2	32.80±0.3	32.15±0.3	9.71±0.8	32.71±0.3	**37.74±0.3**

5. Conclusion

In this paper, we introduced the Topic Model Kernel (TMK), and compare it to other existing kernels in SVM classification tasks for multinomial distribution data types. The significant applications of this work in real-world data are examined on the probabilistic features derived from recent topic models of LDA, HDP, and MC2. These extracted features are more condensed, richer in semantic, and more informative for classification than the raw feature. We confirm that the TMK outperforms other existing kernels on topic modelling feature (drawn from probabilistic assumption). Further, we show that the probabilistic feature extracted from multiple observation model is better than from single observation model. In addition, we demonstrate its comparative performance on the generic types of data (without any assumption of distribution) with the application of digit recognition. Moreover, we investigate that combining raw feature with the extracted feature from probabilistic model would increase the performance. In future work, we would also like to investigate the applicability of the proposed kernel in other distributional-based data and exponential family in general. We believe that the possibility of integrating exponential family in data modelling is still a prolific horizon for our kernel.

References

[1] J Antolín, JC Angulo, and S López-Rosa. Fisher and jensen–shannon divergences: Quantitative comparisons among distributions. application to position and momentum atomic densities. The Journal of chemical physics, **130**:074110, 2009.

[2] Jayanta Basak. A least square kernel machine with box constraints. In Pattern Recognition, 2008. ICPR 2008. 19th International Conference on, pages **1–4**. IEEE, 2008.

[3] D.M. Blei, A.Y. Ng, and M.I. Jordan. Latent Dirichlet allocation. Journal of Machine Learning Research, 3:**993–1022**, 2003. ISSN 1533-7928.

[4] B.E. Boser, I.M. Guyon, and V.N. Vapnik. A training algorithm for optimal margin classifiers. pages **144–152**, 1992.

[5] Sabri Boughorbel, J-P Tarel, and Nozha Boujemaa. Conditionally positive definite kernels for svm based image recognition. In Multimedia and Expo, 2005. ICME 2005. IEEE International Conference on, pages **113–116**. IEEE, 2005.

[6] Antoni B Chan, Nuno Vasconcelos, and Pedro J Moreno. A family of probabilistic kernels based on information divergence. Univ. California, San Diego, CA, Tech. Rep. SVCL-TR-2004-1, 2004.

[7] C.C. Chang and C.J. Lin. Libsvm: a library for support vector machines. ACM Transactions on Intelligent Systems and Technology (TIST), 2(3):**27**, 2011

[8] C. Cortes and V. Vapnik. Support-vector networks. Machine learning, 20(3):**273–297**, 1995.

[9] D.M. Endres and J.E. Schindelin. A new metric for probability distributions. Information Theory, IEEE Transactions on, 49(7):**1858–1860**, 2003.

[10] T.S. Ferguson. A Bayesian analysis of some nonparametric problems. The Annals of Statistics,1(2):209-230,1973.
URLhttp://www.jstor.org/stable/pdfplus/2958008.pdf.

[11] François Fleuret and Hichem Sahbi. Scale-invariance of support vector machines based on the triangular kernel. In 3rd International Workshop on Statistical and Computational Theories of Vision, 2003.

[12] M. Fritz and B. Schiele. Decomposition, discovery and detection of visual categories using topic models. Computer Vision and Pattern Recognition, 2008. IEEE Conference on, pages **1–8**, 2008.

[13] Yoav Goldberg and Michael Elhadad. splitsvm: fast, space-efficient, non-heuristic, polynomial kernel computation for nlp applications. In Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies: Short Papers, pages **237–240**. Association for Computational Linguistics, 2008.

[14] Steve R Gunn. Support vector machines for classification and regression. ISIS technical report, 14, 1998.

[15] Thomas Hofmann. Probabilistic latent semantic indexing. In Proc. ACM SIGIR Int. Conf. on Research and Development in Information Retrieval, pages 50–57, New York, NY, USA, 1999. ACM. ISBN 1-58113-096-1. doi: http://doi.acm. org/10.1145/312624.312649. URL http://www.cs.brown.edu/~th/papers/ Hofmann-SIGIR99.pdf.

[16] Stefan Horbelt, Arrate Munoz, Thierry Blu, and Michael Unser. Spline kernels for continuous-space image processing. In Acoustics, Speech, and Signal Processing, 2000. ICASSP'00. Proceedings. 2000 IEEE International Conference on, volume 6, pages **2191–2194**. IEEE, 2000.

[17] Chih-Wei Hsu, Chih-Chung Chang, Chih-Jen Lin, et al. A practical guide to support vector classification, 2003.

[18] S. Kullback and R.A. Leibler. On information and sufficiency. The Annals of Mathematical Statistics, 22(1):**79–86**, 1951.

[19] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. Proceedings of the IEEE, 86(11): **2278–2324**, 1998.

[20] Pedro J Moreno, Purdy Ho, and Nuno Vasconcelos. A kullback-leibler divergence based kernel for svm classification in multimedia applications. Advances in neural information processing systems, 16:**1385–1393**, 2003.

[21] T.C. Nguyen, D. Phung, S. Gupta, and S. Venkatesh. Extraction of latent patterns and contexts from social honest signals using hierarchical dirichlet processes. 2013 IEEE International Conference on Pervasive Computing and Communications (PerCom 2013), 2013a.

[22] Tien V. Nguyen, D. Phung , and S. Venkatesh. Topic model kernel: An empirical study towards probabilistically reduced features for classification. In International Conference on Neuron Information Processing (ICONIP), 2013b.

[23] V. Nguyen, D. Phung, X. Nguyen, S. Venkatesh, and H. Bui. Bayesian nonparametric multilevel clustering with group-level contexts. In Proc. of International Conference on Machine Learning (ICML), pages 288–296, Beijing, China, 2014a.

[24] Vu Nguyen, Dinh Phung, XuanLong Nguyen, S Venkatesh, and Hung Bui. Supplementary material for Bayesian nonparametric multilevel clustering with contexts. 2014b.

[25] Aude Oliva and Antonio Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. International journal of computer vision, 42(3):145–175, 2001.

[26] D. Phung, X. Nguyen, H. Bui, T.V. Nguyen, and S. Venkatesh. Conditionally dependent Dirichlet processes for modelling naturally correlated data sources. Technical report, Pattern Recognition and Data Analytics, Deakin University, 2012. URL: http://prada-research.net/~dinh/uploads/Main/Publications/ phung_etal_tr12.pdf.

[27] Robert Schlaifer and Howard Raiffa. Applied statistical decision theory. 1961.

[28] J. Shawe-Taylor and N. Cristianini. Kernel methods for pattern analysis. Cambridge Univ Pr, 2004.

[29] J. Shlens. Notes on kullback-leibler divergence and likelihood theory. System Neurobiology Laboratory, Salk Institute for Biological Studies, California, 2007.

[30] Y.W. Teh, M.I. Jordan, M.J. Beal, and D.M. Blei. Hierarchical Dirichlet processes. Journal of the American Statistical Association, 101(476):**1566–1581**, 2006.

[31] Flemming Topsoe. Jenson-shannon divergence and norm-based measures of discrimination and variation. Preprint, 2003.

[32] Christian Wartena and Rogier Brussee. Topic detection by clustering keywords. In Database and Expert Systems Application, 2008. DEXA'08. 19th International Workshop on, pages **54–58**. IEEE, 2008.

Received January 15, 2015; accepted March 10, 2015.

Vu Nguyen
Center for Pattern Recongnition and Data Analytics (PRaDA)
Deakin Unversity, Melbourne, Australia
tvnguye@deakin.edu.au