DEAKIN UNIVERSITY AUSTRALIA

Topic Model Kernel: An Empirical Study Towards Probabilistically Reduced Features For Classification

Tien-Vu Nguyen, Dinh Phung, Svetha Venkatesh

Center for Pattern Recognition and Data Analytics (PRaDA)

Deakin University, Australia

DEAKIN UNIVERSITY AUSTRALIA

Content

1. Introduction
2. Support Vector Machine
3. Topic Model Feature
4. Topic Model Kernel for Classification
5. Experiments
6. Question & Answer

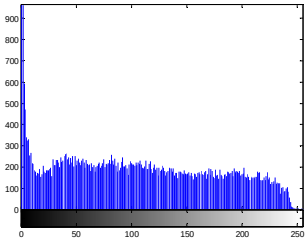
1. Introduction

Classification.

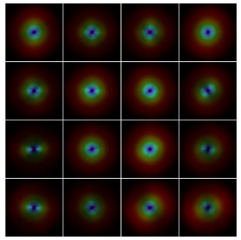
- The low-level features may not able to semantically represent the data well.

Low-level representation
(high-dimensional feature)

High-level representation
(low-dimensional feature)



Color Histogram



GIST

RBM
Neuron Network
Topic Models
…

1. Introduction

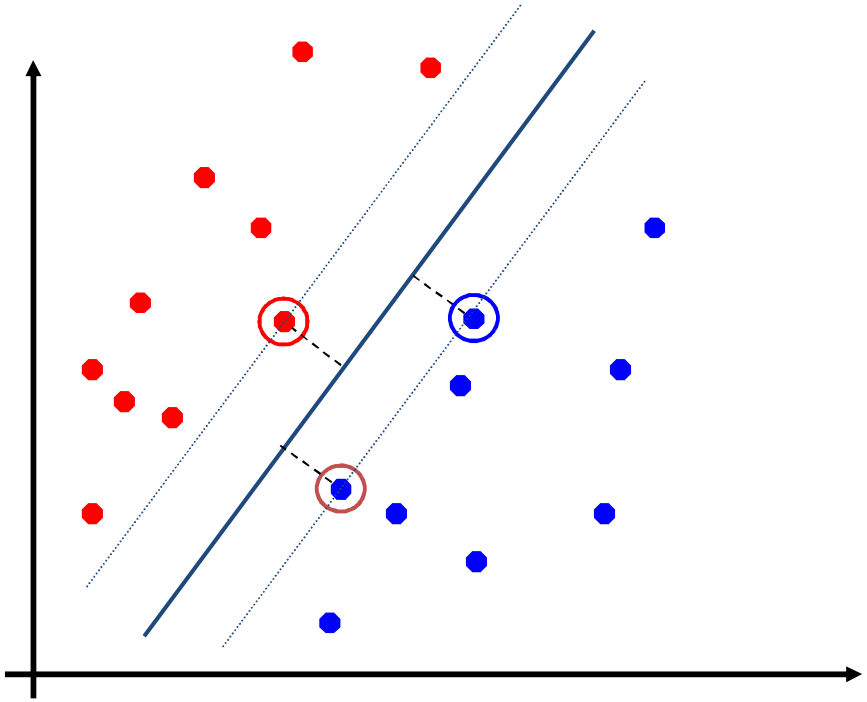
Classification.

- The low-level features may not able to semantically represent the data well.
- The integrated classification framework:
  - (1) discovering the “high-level” features (e.g. topic models)
  - (2) do classification with supervised learning method (e.g. SVM, ...).
- We propose the Topic Model Kernel (TMK) for SVM classification with the topic model feature.

DEAKIN UNIVERSITY AUSTRALIA

2. Support Vector Machine

- Supervised learning technique.
- Finding the line that maximizes the margin between classes.



Machine Learning Group
University of Texas Austin

3. Topic Model

discovering the main “topics” from collections of “documents”

“documents”

Applications

- Summarization
- Classification
- Information retrieval

3. Example of Topics learned by Topic Model

sky sea water rocks sand beach

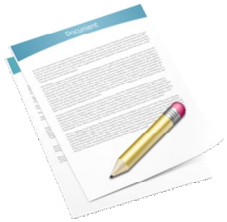
door window balcony

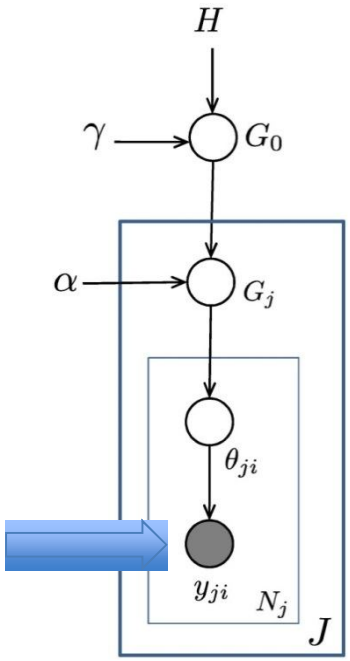
person walking building poster plants

Topic Models

Parametric: Latent Dirichlet Allocation (LDA), TOT, sLDA…
Nonparametric: Hierarchical Dirichlet Process (HDP), CDDP…

Hierarchical Dirichlet Processes

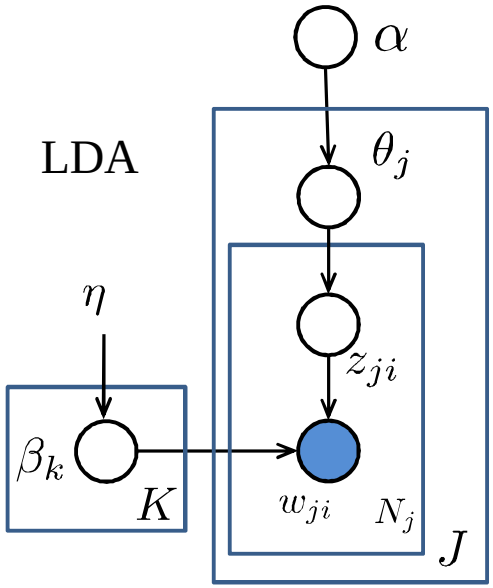




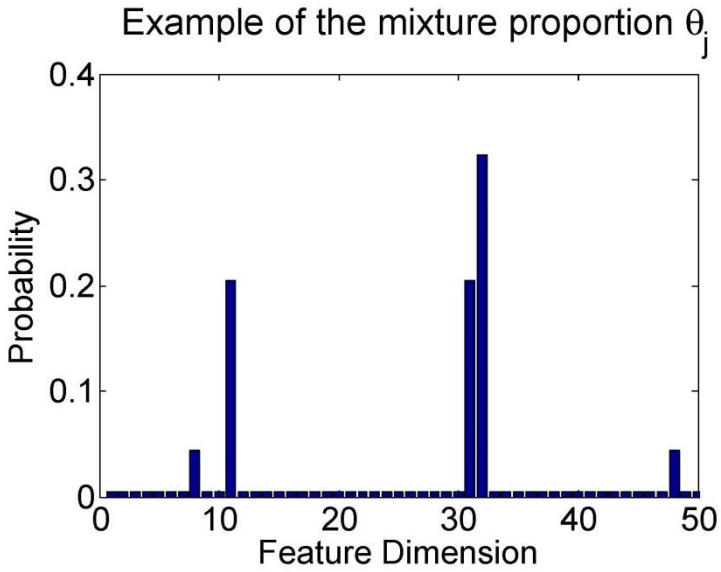
Dirichlet Process

Automatically identify number of topic

Topic Model Feature



Reduced from original dimension of 65,483 to 50



Topic model feature, high-level representation, follows Multinomial distribution.

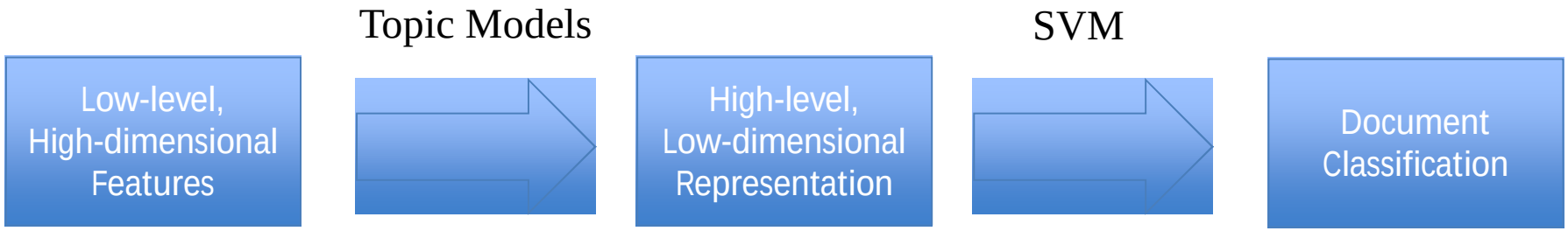
Topic model features are condensed in low dimension,
they are more semantically informative and discriminative for classification

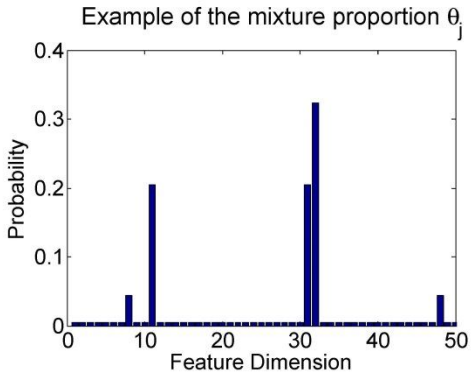
DEAKIN UNIVERSITY AUSTRALIA

Content

1. Introduction
2. Support Vector Machine
3. Topic Model Feature
4. Topic Model Kernel for Classification
5. Experiments
6. Question & Answer

Classification with Topic Model Feature





DEAKIN UNIVERSITY AUSTRALIA

4. Topic Model Kernel

P and Q are two probabilistic distributions (e.g. Multinomial, Gaussian).

Kullback–Leibler (KL) Divergence

$$D_{KL}\left(P \parallel Q\right) = \sum_{i} P(i) \ln \frac{P(i)}{Q(i)}$$

Is not a metric

Jensen-Shanon (JS) Divergence

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

$$M = \tfrac{1}{2}\left(P + Q\right)$$

JS squared root is a metric for two probability distributions. (D. Endres, J. Schindelin, 2003)

It is negative definite on $R_+ \times R_+$ (Topsoe, 2003)

DEAKIN UNIVERSITY AUSTRALIA

4. Topic Model Kernel

X and Y are two (non-negative) feature vectors.(e.g Topic Model Feature)

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

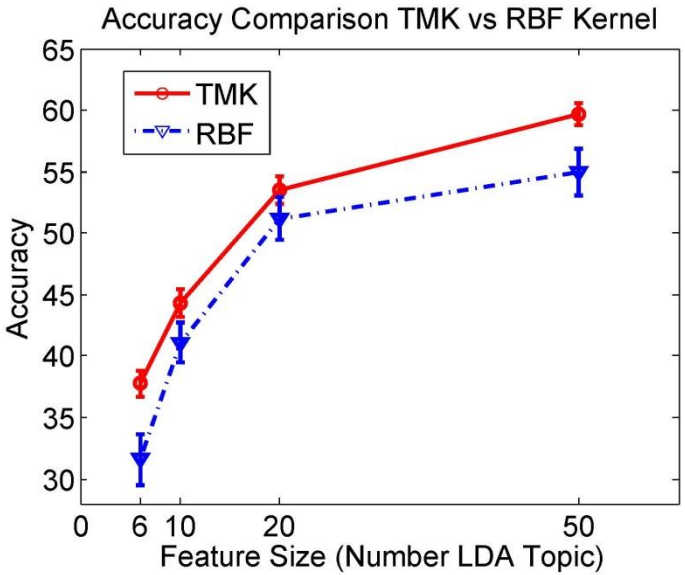
$$= \exp\left\{-\frac{1}{\sigma^2} \times \left[\frac{1}{2}\sum_i X(i) \ln \frac{X(i)}{M(i)} + \frac{1}{2}\sum_i Y(i) \ln \frac{Y(i)}{M(i)}\right]\right\}$$

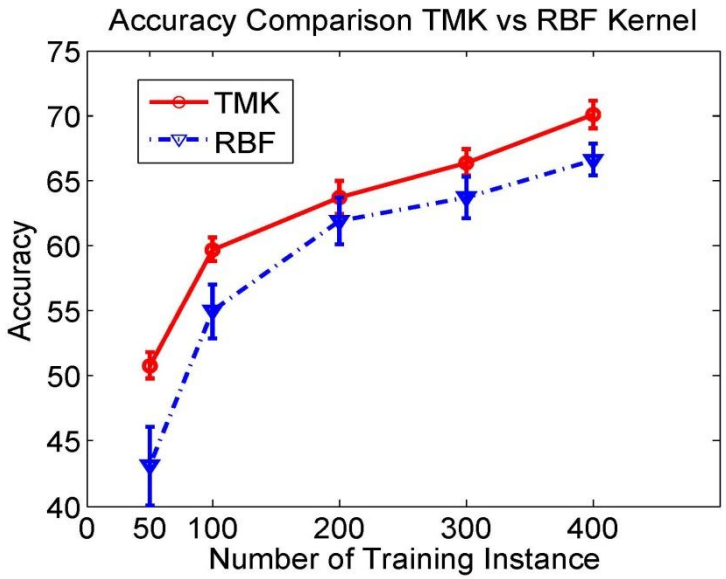
It satisfied a positive semi definite condition to be a kernel.

5. Experimental Results

Livejournal corpus: classify communities

8758 posts in 65,483-dimension raw feature
using LDA…
8758 posts in k-dimension (e.g 6,10,20,50…)





DEAKIN UNIVERSITY AUSTRALIA

5. Experimental Results on Topic Model Features

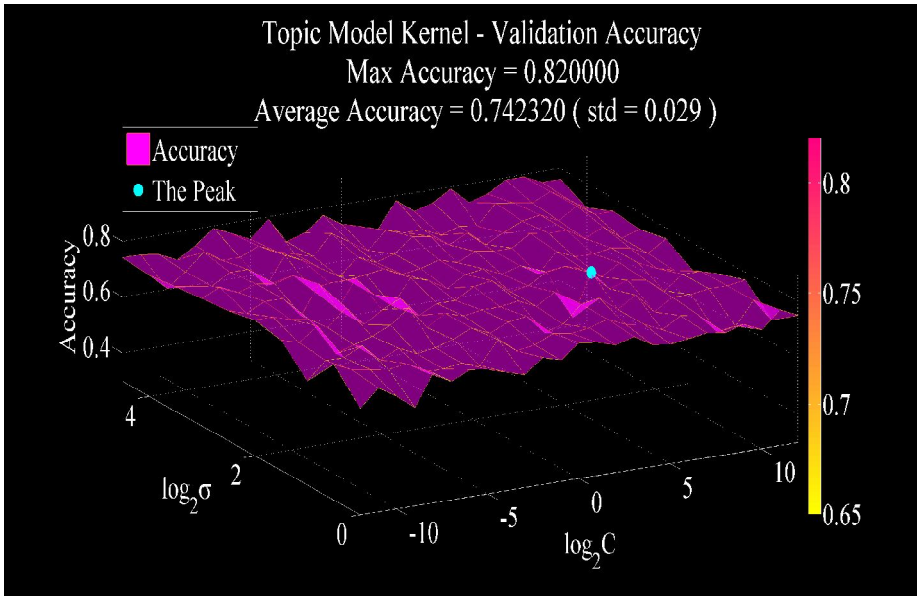
Dataset	TMK	RBF	Linear	Polynomial	Sigmoid
Livejournal	**58.1 (2.1)**	54.9 (4.9)	54.4 (5.2)	52.6 (6.6)	51.8 (5.1)
Reuter21578	**81.3 (0.2)**	79.0 (0.5)	78.2 (0.1)	77.9 (0.1)	77.4 (0.4)
LabelMe	**72.3 (1.9)**	70.8 (1.9)	71.5 (1.8)	62.7 (4.2)	69.8 (1.6)

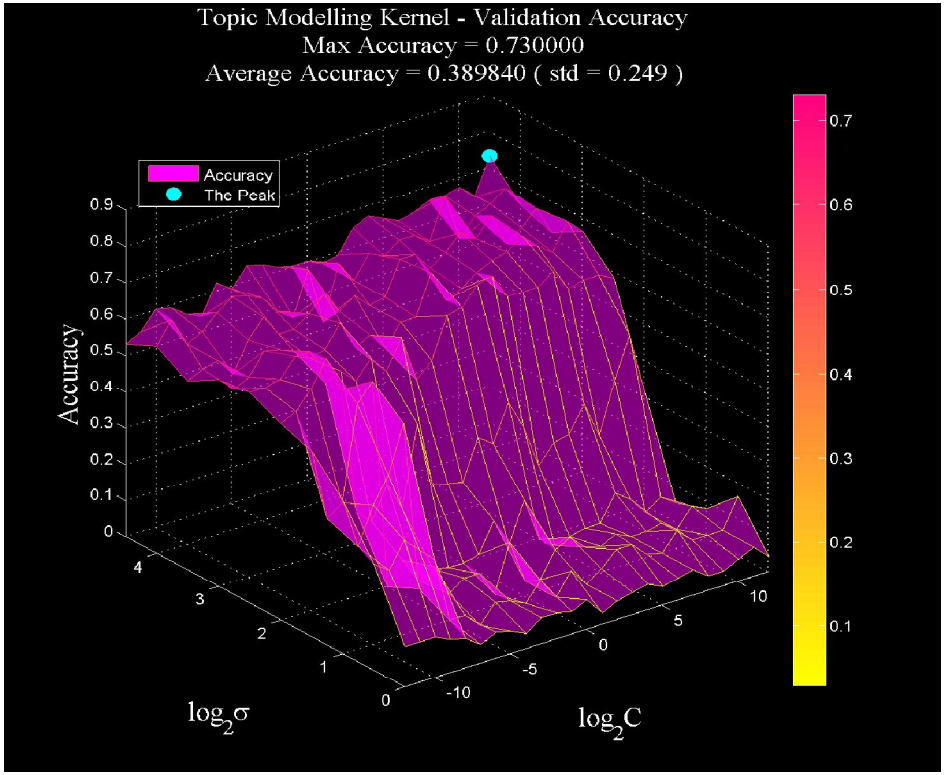
DEAKIN UNIVERSITY AUSTRALIA

5. Experimental Results on Generic Features

Dataset	TMK	RBF	Linear	Polynomial	Sigmoid
MNIST (at default parameter)	88.4(0.9)	82.2(1.9)	**91.3(0.5)**	83.7(2.6)	79.6(3.4)
MNIST (at optimal parameter by cross-validation)	**90.8(2.25)**	89.3(2.5)	89.8(2.7)	88.7(2.9)	85.8(2.7)

5. Experimental Results





Excellent performance
Robust on topic model features
(LabelMe)

Very good performance
but not stable on generic feature
(MNIST)

Summary

- Document classification with two steps:
  1. Topic Models for extracting low-dimentional feature.
  2. Do classification by SVM with Topic Model Kernel.

- TMK is superior on Topic Model feature, comparable on other types of feature.

TMK Matlab code is available at author website.
prada-research.net/~tienvu/code

TMK Matlab code is available:

prada-research.net/~tienvu/code