Applications of Information Bottleneck

Document Classification

Mika Pollari 9.3.2004

Outline

- Introduction: Information Bottleneck (IB) framework in document clustering
- Theory of IB method
- Sequential IB *(sIB)* approach to unsupervised Document Classification
- IB approach to feature selection for text Categorization
- Conclusions

Motivation

- The Basic idea: Find a clusters for a collection of documents that are correlated with (true) topics of the documents
- Document classification is needed for large document collections, applications:

  1 Information retrieval form large collection

  2 Navigating and browsing large collections

Introduction: IB framework in document classification

Distance/Distortion:

Clustering algorithms are based on distance/distortion measure. In document classification a natural measure of similarity between two documents is based on their word conditional distributions ***p(y|x)*** ➔ *Documents with similar conditional word probabilities should belong to same cluster*

Selecting 'right' distance/distortion measure:

How to select 'right' distance/distortion measure between distribution? Arbitrary selection?? IB approach provides an answer.

Introduction: IB framework (cont.)

IB approach:

Given the joint distribution *p(X,Y)* of documents (*X*) and vocabulary *(Y)*, look for a compact representation (*T*) of *X*, which preserves information as much as possible about variable *Y*

Mutual Information:

Is a natural measure of the information that variable *X* contains about variable *Y*

$$I(X;Y) = \sum_{x \in X} \sum_{y \in Y} p(x) p(y \mid x) \log \frac{p(y \mid x)}{p(y)}$$

Theory of IB method

Find a partition ***T(X)*** which maximizes a score function ***F(T).*** Score function is defined through another variable *Y*

IB method maximize mutual information I(T;Y) under constrain on I(T;X)

➔ $$F(T) = \max_{p(T|X)}(I(T;Y) - bI(T;X))$$

Theory of IB method (cont.)

The solution satisfied following equations (1-3)

$$\begin{cases} p(t \mid x) = \dfrac{p(t)}{Z(b, x)} \exp(-b D_{KL}(p(y \mid x) \| p(y \mid t)) & (1) \\ p(y \mid t) = \dfrac{1}{p(t)} \sum_{x \in X} p(t \mid x) p(x) p(y \mid x) & (2) \\ p(t) = \sum_{x \in X} p(t \mid x) p(x) & (3) \end{cases}$$

Solution can be obtained by starting from arbitary solution and iterating the equations. For any value of *b* procdure will converge. **Note!!** Different values of *b* corresponds to different distributional resolutions (number of clusters)

Sequantial IB clustering

IB framework is used to classify unlabeled documents: X = document; Y=vocabulary; T=clusters

Preprosessing: Building vocabulary: ignore non alpha-numeric characters, unite digits to one symbol, remove words that occure only once etc...

Sequantial IB clustering (cont.)

Nuber of clusters is fixed (K clusters)

➔ $F(T)=I(T;Y)$

➔ $d(x,t)=(p(x)+p(t))JS(p(y|x),p(y|t))$

Pseudo code for sIB:

Input:

|**X**| documents; Parameter ***K***

Output:

A partition ***T*** of ***X*** into ***K*** clusters

Main loop:

T ← random partition of X

do

for j=1,…,|X|

draw x_j out of $t(x_j)$

$t_{new}(x_j) = \arg\min d_F(x_j, t)$

merge x_j into t_{new}

end

until(converge)

Relation to formal solution

Every partition defines a hard propability $p(t|x)$ which in turn defines propabilities $p(t)$ and $p(y|t)$ through **Eqs. 1-3**

Assume that t_{new} differs from t then $F(T_{new})>F(T)$ because $t_{new}=arg\ min_{t \in T}\ d_F(x,t)$. The convergence (to local optima) is guaranteed because F(T)=I(T;Y) has an upper bound I(X;Y)

Improvements to *sIB* method

Restart the algorithm #n times with different initial partitions and select the best result to avoid local optima:

$$T = \arg\max_{Ti} F(Ti)$$

Estimate *d(x,t(x))* and use top *r%*

- if document in top r% ---> label document
- otherwise document is unlabeled (higher precision is gained)

Advantages of *sIB*

The time and space complexity are improved from standard IB (down-to-top)

sIB is guaranteed to converge to a local maximum

Better classification results

IB framework in feature selection

Use IB framework to calculate compact and efficient representation of documents (word clusters)

X = vocabulary (from training set)

T = word clusters

Y = class labels (from training set)

Features (word clusters) are given for SVM classifier

IB framework in feature selection (cont.)

Another way to select "features" for classifier is to use bag-of-words (BOW) combined with MI feature selection:

For each category, *c*, select most discriminative words of the category by selecting top ***k*** words according to:

$$I(X_C; X_W)$$

IB framework in feature selection (cont.)

The score function to maximized:

$$F(T) = \max_{p(T|X)}(I(T;Y) - bI(T;X))$$

The solution is obtained from Eqs(1-3)

Pseudo code for algorithm:

- Input:
  - p(X,Y), K (number of clusters), values for beta
- Output:
  - cluster centroids p(Y,T) and assignment probabilities p(T|X)

$\beta \leftarrow \beta_{min}$ r$\leftarrow$1 (number of centroids)

repeat

    compute *p(T|X);p(T);p(Y|T) eqs. 1-3*

    repeat

        p_{old}(T|X)$\leftarrow$p(T|X)

        compute *p(T|X);p(T);p(Y|T) eqs. 1-3*

    until for each x: $||p_{old}(T|x)-p(T|x)||$<convergence

    for all centroids i,j and $||p(Y|ti)-P(Y|tj)||$<merge

        merge ti and tj; r$\leftarrow$r-1; p(ti|X)=p(ti|X)+p(tj|X)

    end for-loop

    for all centroids i

        create tr+1 $||p(Y|tr+1)-P(Y|t)||$<merge

        p(ti|X)$\leftarrow$0.5p(ti|X), p(tr+1|X)$\leftarrow$0.5p(ti|X)

    end for; r$\leftarrow$2r β=sβ

until (r>K or max $\beta > \beta_{max}$)

IB framework in feature selection (cont.)

- Output of the algorithm (cluster centroids p(Y,T)) are the input for the SVM classifier
- Test results show that IB feature selection is equally good as BOW+MI selection
- The advantage of IB feature selection is that algorithm discovers global features where as BOW+MI selection must be done separately for each document category

Conclusions

The IB principle is based on "distributional clustering" under relevance variable Y

The score function to maximized:

$$F(T) = \max_{p(T|X)} (I(T;Y) - bI(T;X))$$

The solution satisfied following equations:

$$\left\{ \begin{array}{c} p(t \mid x) = \dfrac{p(t)}{Z(b, x)} \exp(-b D_{KL}(p(y \mid x) \| p(y \mid t)) \\ p(y \mid t) = \dfrac{1}{p(t)} \sum_{x \in X} p(t \mid x) p(x) p(y \mid x) \\ p(t) = \sum_{x \in X} p(t \mid x) p(x) \end{array} \right\}$$

Conclusions (cont.)

IB framework can be efficiently used in documents classification or in feature selection for another classifier

IB framework is very flexible different clustering approaches (sequential, top-down, bottom-up) are possible

Convergence to local minimum is guaranteed??