ON HORIZONTAL AND VERTICAL SEPARATION IN HIERARCHICAL TEXT CLASSIFICATION

MOSTAFA DEHGHANI

University of Amsterdam
HIERARCHICAL STRUCTURE

- An effective and common way of representing information:
  - Determines relationships in the data at different levels of resolution
One the key concepts of information retrieval:

Especially when the task is making a boolean decision:
SEPARATION IN HIERARCHIES

- **Two types of dependencies in the hierarchies:**
  - **Horizontal dependency:**
    - Relations of entities in the same layer
  - **Vertical dependency:**
    - Relations between ancestors and descendants
ON HORIZONTAL AND VERTICAL SEPARATION IN HIERARCHICAL TEXT CLASSIFICATION

HORIZONTAL SEPARATION

- Typical problem of classification
HORIZONTAL SEPARATION

- Typical problem of classification
HORIZONTAL SEPARATION

- Typical problem of classification
VERTICAL SEPARATION

- How vertical separation is beneficial property?
VERTICAL SEPARATION

- How vertical separation is beneficial property?

  - Let’s see an example!

```
All  ❯ Human
  ❯ Roles  ❯ Former US President  ❯  …
  ❯ Person  ❯ Obama  ❯ Bush  ❯ Clinton  ❯  …
  ❯ Properties  ❯  …  ❯  …  ❯  …  ❯  …
```

m__dehghani
How vertical separation is beneficial property?

Let’s see an example!
VERTICAL SEPARATION

- How vertical separation is beneficial property?

- Let’s see an example!
Probability Ranking Principle (PRP) for binary classification: [Lewis, 1995]

For a given set of items presented to a binary classification system, there exists a classification of the items such that the probability of class membership for all items assigned to the class is greater than or equal to the probability of class membership for all items not assigned to the class, and the classification has optimal expected effectiveness.

Probability Threshold Principle (PTP):

For a given effectiveness measure, there exists a threshold \( p \), \( 0 < p < 1 \), if all and only those items with the probability of class membership greater than \( p \) are assigned to the class, the expected effectiveness of the classification will be the best possible for that set of items.

“separability in the score space”

ON HORIZONTAL AND VERTICAL \textbf{SEPARATION} IN HIERARCHICAL TEXT \textbf{CLASSIFICATION}

FROM PROBABILITY RANKING PRINCIPLE TO STRONG SEPARATION PRINCIPLE

“separability in the \textbf{Feature} space”
Definition 1. The model of an entity is epistemologically “separable” if, and only if, it has unique, non-overlapping features that distinguish it from other models.

Strong Separation Principle (SSP):

For a given set of items presented to a classification system, for each class there exists at least one feature $\delta$ in the representation of items, and a threshold $\tau$, such that for any set of items, if all and only those items with $\delta > \tau$ are assigned to the class, the classification will have the optimal possible performance for that set of items in terms of a given effectiveness measure.
Definition 2. The model of an entity in the hierarchy is "horizontally separable" if, and only if, it is separable compared to other entities in the same layer, with the same abstraction level.

\[ p(t|\theta_c) = \lambda p(t|\theta_c^{hs}) + (1 - \lambda)p(t|\theta_c^g) \]
Definition 2. The model of an entity in the hierarchy is "horizontally separable" if, and only if, it is separable compared to other entities in the same layer, with the same abstraction level.

\[
p(t|\theta_c) = \lambda p(t|\theta_c^{hs}) + (1 - \lambda) p(t|\theta_c^g)
\]

\[
\approx \lambda p(t|\theta_c^{hs}) + (1 - \lambda) p(t|\theta_{all})
\]

\[
p(t|\theta_{all}) = \frac{tf(t, all)}{\sum_{t'} tf(t', all)} = \frac{\sum_{c \in all} \sum_{d \in c} tf(t, d)}{\sum_{c \in all} \sum_{d \in c} \sum_{t' \in d} tf(t', d)}
\]
Definition 2. The model of an entity in the hierarchy is "horizontally separable" if, and only if, it is separable compared to other entities in the same layer, with the same abstraction level.

\[
p(t | \theta_c) = \lambda p(t | \theta_{hs}^c) + (1 - \lambda) p(t | \theta_{all}^c)
\]

\[
\approx \lambda p(t | \theta_{hs}^c) + (1 - \lambda) p(t | \theta_{all}^c)
\]

\[
p(t | \theta_{all}) = \frac{\sum_{t \in c} t f(t, all)}{\sum_{t' \in \text{all}} t f(t', all)} = \frac{\sum_{c \in \text{all}} \sum_{d \in c} t f(t, d)}{\sum_{c \in \text{all}} \sum_{d \in c} \sum_{t' \in d} t f(t', d)}
\]

\[
\log p(t | \theta_{hs}^c) = \sum_{t \in c} t f(t, c) \log \left( \lambda p(t | \theta_{hs}^c) + (1 - \lambda) p(t | \theta_{all}^c) \right)
\]

E - Step:

\[
e_t = t f(t | c) \cdot \frac{\lambda p(t | \theta_{hs}^c)}{\lambda p(x | \theta_{hs}^c) + (1 - \lambda) p(x | \theta_{all})}
\]

M - Step:

\[
p(x | \theta_{hs}^c) = \frac{e_t}{\sum_{t' \in \mathcal{V}} e_{t'}}
\]

i.e. normalizing the model
Definition 3. The model of an entity in the hierarchy is “vertically separable” if, and only if, it is separable compared to other entities in the other layers, with different abstraction levels.

\[
p(t|\theta_c) = \lambda p(t|\theta_{c}^{os}) + (1 - \lambda)p(t|\theta_{d}^{s})
\]

\[
p(t|\theta_{d}^{s}) \leftarrow \text{normalized} \sum_{d_i \in C} \left( p(t|\theta_{d_i}) \prod_{d_j \in C \setminus d_i} (1 - p(t|\theta_{d_j})) \right)
\]
Definition 3. The model of an entity in the hierarchy is “vertically separable” if, and only if, it is separable compared to other entities in the other layers, with different abstraction levels.

\[
p(t|\theta_c) = \lambda p(t|\theta_c^{\text{vs}}) + (1 - \lambda)p(t|\theta_d^s)
\]

\[
p(t|\theta_d^s) \leftarrow \frac{\text{normalized}}{\sum_{d_i \in c} \left( p(t|\theta_{d_i}) \prod_{d_j \in c, j \neq i} (1 - p(t|\theta_{d_j})) \right)}
\]

E-Step: 

M-Step: 

...
Definition 4. The model of an entity in the hierarchy is “two-dimensionally separable” if, and only if, it is both horizontally and vertically separable at the same time.
Definition 4. The model of an entity in the hierarchy is "two-dimensionally separable" if, and only if, it is both horizontally and vertically separable at the same time.

\[
p(t|\theta_e) = \lambda_1 p(t|\theta_e^{specific}) + \lambda_2 p(t|\theta_e^{hvs}) + \lambda_1 p(t|\theta_e^{general})
\]
Definition 4. The model of an entity in the hierarchy is “two-dimensionally separable” if, and only if, it is both horizontally and vertically separable at the same time.

\[
p(t|\theta_e) = \lambda_1 p(t|\theta_e^{specific}) + \lambda_2 p(t|\theta_e^{hvs}) + \lambda_1 p(t|\theta_e^{general})
\]
Definition 4. The model of an entity in the hierarchy is "two-dimensionally separable" if, and only if, it is both horizontally and vertically separable at the same time.

\[
p(t|\theta_e) = \lambda_1 p(t|\theta_e^{specific}) + \lambda_2 p(t|\theta_e^{hvs}) + \lambda_1 p(t|\theta_e^{general})
\]
Definition 4. The model of an entity in the hierarchy is “two-dimensionally separable” if, and only if, it is both horizontally and vertically separable at the same time.
Definition 4. The model of an entity in the hierarchy is “two-dimensionally separable” if, and only if, it is both horizontally and vertically separable at the same time.

\[
p(t|\theta_e) = \lambda_1 p(t|\theta^{specific}_e) + \lambda_2 p(t|\theta^{hvs}_e) + \lambda_1 p(t|\theta^{general}_e)
\]
HIERARCHICAL SIGNIFICANT WORDS LANGUAGE MODELS

Algorithm Estimating Hierarchical Significant Words Language Models

1: procedure ESTIMATEHSLM
2: Initialization:
3: for all entity e in the hierarchy do
4: \[ \theta_e \leftarrow \text{standard estimation for } e \text{ using MLE} \]
5: end for
6: Specification
7: Generalization
8: until models do not change significantly anymore
9: end procedure

Algorithm Modified Model Parsimonization

1: procedure PARSIMONIZE(e,B)
2: for all term t in the vocabulary do
3: \[ P(t|\theta_B) \leftarrow \frac{\sum_{b_i \in B} (P(t|\theta_{b_i}) \prod_{b_j \in B(1 - P(t|\theta_{b_j}))})}{j \neq i} \]
4: repeat
5: E-Step: \[ P[t \in V] \leftarrow P(t|\theta_e) \cdot \frac{\alpha P(t|\theta_e)}{\alpha P(t|\theta_e) + (1 - \alpha) P(t|\theta_B)} \]
6: M-Step: \[ P(t|\theta_e) \leftarrow \frac{P[t \in V]}{\sum_{t' \in V} P(t' \in V)} \]
7: until \[ \bar{\theta}_t \] becomes stable
8: end for
9: end procedure

Specification Stage

1: procedure SPECIFICATION
2: Queue \leftarrow all entities in breadth first order
3: while Queue is not empty do
4: \[ e \leftarrow \text{Queue.pop()} \]
5: \[ l \leftarrow e.\text{Depth()} \]
6: while \[ l > 0 \] do
7: \[ A \leftarrow e.\text{GETANCESTOR}(l) \]
8: PARSIMONIZE(e,A)
9: \[ l \leftarrow l - 1 \]
10: end while
11: end while
12: end procedure

Generalization Stage

1: procedure GENERALIZATION
2: Stack \leftarrow all entities in breadth first order
3: while Stack is not empty do
4: \[ e \leftarrow \text{Stack.pop()} \]
5: \[ l \leftarrow e.\text{Height()} \]
6: while \[ l > 0 \] do
7: \[ D \leftarrow e.\text{GETDECEDENTS}(l) \]
8: PARSIMONIZE(e,D)
9: \[ l \leftarrow l - 1 \]
10: end while
11: end while
12: end procedure
EXPERIMENTS:

- Dataset:
  - Dutch parliamentary data

- Horizontal Separability:
  - Probability distribution over terms based on hierarchical significant words language models in status layer and party layer.

(a) HSWLM in the status layer

(b) HSWLM in the party layer
ON HORIZONTAL AND VERTICAL SEPARATION IN HIERARCHICAL TEXT CLASSIFICATION

EXPERIMENTS:

- **Horizontal Separability:**

  probability distribution over terms based on hierarchical significant words language models in party layer.

(a) HSWLM of two parties in different statuses: Christian Democratic Appeal (CDA) and Labour Party (PvdA)

(b) HSWLM of two parties in opposition: Party for Freedom (PVV) and Christian Democratic Appeal (CDA)

(c) HSWLM of two parties in government: People’s Party for Freedom (VVD) and Labour Party (PvdA)
ON HORIZONTAL AND VERTICAL SEPARATION IN HIERARCHICAL TEXT CLASSIFICATION

EXPERIMENTS:

- **Vertical Separability:**

  - probability distribution over terms in different layers based on hierarchical significant words language models in complete paths from the root to the terminal entities in the hierarchy

  \[\text{(a) HSWLM of S. van Haersma Buma (as the member of parliament - Leader of CDA), Christian Democratic Appeal (as the party), Opposition (as the status), and the Parliament}}\]

  \[\text{(b) HSWLM of D. Samson (as the member of parliament - Leader of PvdA), Labour Party (as the party), Government (as the status), and the Parliament}}\]
ON HORIZONTAL AND VER TICAL SEPARATION IN HIERARCHICAL TEXT CLASSIFICATION

EXPERIMENTS: SEPARABILITY FOR TRANSFERABILITY

Table 1: Results of party classification task in terms of macro-average accuracy. We have conducted paired t-test to investigate statistical significance of the improvements of the best method over the second best method, in the corresponding experiments. Improvements that are annotated with * are statistically significant with p-value < 0.005.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Train</td>
<td>SVM</td>
<td>SVMIG</td>
<td>SVMHSLM</td>
<td>SVM</td>
<td>SVMIG</td>
<td>SVMHSLM</td>
</tr>
<tr>
<td>2010-2012</td>
<td>40.90</td>
<td>35.57</td>
<td>43.11*</td>
<td>34.12</td>
<td>41.83</td>
<td>40.02*</td>
</tr>
<tr>
<td>2012-2014</td>
<td>30.51</td>
<td>44.96</td>
<td>30.38</td>
<td>47.18</td>
<td>39.11*</td>
<td>47.28</td>
</tr>
</tbody>
</table>
As an extrinsic evaluation of the estimated models, we investigate the question: "How effective are hierarchical language models?" Hierarchical models, especially those based on the significant words hierarchy, provide robust models by taking out objects in the hierarchy as it is explained in Section 3.

Lemmatization is done on the data and also stop words and common words are not included. We then investigate our second research question: "How effective are hierarchical language models?" Hierarchical models, especially those based on the significant words hierarchy, provide robust models by taking out objects in the hierarchy as it is explained in Section 3.

As expected, the diversity of features for different parties in a single document which textually represent them: first over parties, and then parties over period. Figure 3 shows the diversity of distributions in each of the three cases for each of the three weighting methods.

Figure 10 shows the average diversity of distributions in each of the cases for each of the three weighting methods. The results indicate that the diversity of features for different parties in a single document which textually represent them: first over parties, and then parties over period. Figure 3 shows the diversity of distributions in each of the three cases for each of the three weighting methods.

Table 1: Results of party classification task in terms of macro-average accuracy. We have conducted paired t-test to investigate statistical significance of the improvements of the best method over the second best method, in the corresponding experiments. Improvements that are annotated with * are statistically significant with p-value < 0.005.

<table>
<thead>
<tr>
<th>Test</th>
<th>SVM</th>
<th>SVM_MIG</th>
<th>SVM_HSWLM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Train</td>
<td>40.90</td>
<td>35.57</td>
<td>43.11*</td>
</tr>
<tr>
<td></td>
<td>30.51</td>
<td>44.96</td>
<td>30.38</td>
</tr>
</tbody>
</table>

- Average diversity of the representation of features of CDA and PvdA in different situations

In this paper, we investigated the separation property in hierarchical text classification. We compared the performance of HSWLM with SVM and SVM_MIG using paired t-test to investigate statistical significance of the improvements of the best method over the second best method, in the corresponding experiments. Improvements that are annotated with * are statistically significant with p-value < 0.005.

<table>
<thead>
<tr>
<th>JS Divergence</th>
</tr>
</thead>
<tbody>
<tr>
<td>Different Parties</td>
</tr>
<tr>
<td>Same Period</td>
</tr>
<tr>
<td>Different Periods</td>
</tr>
<tr>
<td>Different Periods</td>
</tr>
</tbody>
</table>

ON HORIZONTAL AND VERTICAL SEPARATION IN HIERARCHICAL TEXT CLASSIFICATION
CONCLUSION

- We had three main Research Questions:
  
  - **RQ1:** “What makes separability in the feature space a desirable property for classifiers?”
    
    We have introduced “Strong Separation Principle” for optimizing expected effectiveness of classifiers
  
  - **RQ2:** “How can we estimate horizontally and vertically separable language models for the hierarchical entities?”
    
    We proposed Hierarchical Significant Words Language Models (HSLWLM) possessing two-dimensional separability
  
  - **RQ3:** “How separability improves transferability?”
    
    We demonstrated that separability makes the model more robust and transferable over time by filtering out non-essential non-stable terms.
THANK YOU!