Novel Properties and Well-Tried Performance of EM-Based Multivariate Clustering

Detlef Prescher

Language Technology Lab
Stuhlsatzenhausweg 3, D-66123 Saarbrücken, Germany
prescher@dfki.de

September 5, 2001

Overview

- Introduction
- Theory
  - Probabilistic Multivariate Clustering
  - Simplified EM-Based Re-Estimation Formulas
  - Invariance Property
- Experiments
  - Pruning
  - Smoothing
- Well-Tried Performance
- Discussion and Conclusion

Introduction

- **Aim:** Unsupervised induction of probabilistic classes for arbitrary linguistic data, e.g. semantic labels for verb-argument pairs or phonological labels for onset-nucleus-coda triples, to
  - automatically reveal the **hidden structural information** in a given linguistic corpus,
  - obtain a **smooth model** on both the data types **in** the corpus and the reasonable data types **not in** the corpus.

- **Constraints:** In detecting and generalizing information, induced probabilistic clustering models should
  - preserve the main characteristics of a given **real-world corpus**, and
  - have a good performance when utilized for **real-world applications**.

Probabilistic Multivariate Clustering

The probability of a d-dimensional data type $y = (y_1, .., y_d) \in \mathcal{Y}_1 \times \ldots \times \mathcal{Y}_d$ is defined as:

$$p(y) = \sum_{c \in C} p(c, y) = \sum_{c \in C} p(c) \cdot p(y|c) = \sum_{c \in C} p(c) \prod_{i=1}^{d} p(y_i|c) \ .$$

The **independence assumption**

$$p(y|c) = \prod_{i=1}^{d} p(y_i|c)$$

for each class c makes clustering feasible since an independence assumption is not adequate for the whole sample, and therefore **forces the sample to break into clusters**.

Multivariate Clustering via Regular Grammars

Given the simple **probabilistic regular grammar**

$$\begin{array}{ll} S \rightarrow Y_1^c \ldots Y_i^c \ldots Y_d^c & \quad p(c) \\ Y_i^c \rightarrow y_i & \quad p(y_i|c) \end{array}$$

with starting symbol S, non-terminals Y_i^c (with $c \in C$ and $i = 1 \ldots d$) and terminals $y_i \in \mathcal{Y}_i$, a "sentence" $y_1 \ldots y_i \ldots y_d$ has exactly $|C|$ "syntax trees":

$$\begin{array}{ccccc} & & S & & \\ Y_1^c & \ldots & Y_i^c & \ldots & Y_d^c \\ y_1 & \ldots & y_i & \ldots & y_d \end{array}$$

Each "syntax tree" has the probability of $p(c) \prod_{i=1}^{d} p(y_i|c)$ yielding a "sentence probability" of $\sum_{c \in C} p(c) \prod_{i=1}^{d} p(y_i|c)$ which is equal to the probabilty of the type $y_1 \ldots y_d$ given the multivariate clustering model.

EM-Based Multivariate Clustering

Using the well-known EM-based re-estimation formulas for probabilistic context-free grammars, simple **EM-based re-estimation formulas** for probabilistic multivariate clustering can be derived:

$$\begin{aligned}\hat{p}(c) &= |f|^{-1} \cdot |f_c|, \\ \hat{p}(y_i|c) &= |f_c|^{-1} \cdot f_c(y_i),\end{aligned}$$

where the total frequencies $|f| := \sum_y f(y)$ and $|f_c| := \sum_y f_c(y)$ are computed using the sample frequencies $f(y)$ and **class-based frequencies**

$$f_c(y) = f(y) \cdot p(c|y),$$

and $f_c(y_i) := \sum_{y \in \mathcal{H}(y_i)} f_c(y)$ is a **marginal class-based frequency** characterized by the **hyper plane** $\mathcal{H}(y_i)$ defined as the set of data types y with fixed coordinate y_i.

Toy-Example: EM-Based 2-dimensional Clustering

Training corpus: given via the so-called **empirical distribution**

$$\tilde{p}(y_1, y_2) := |f|^{-1} \cdot f(y_1, y_2).$$

$\tilde{p}(y_1, y_2)$	smile	laugh	increase	fall
man	0.2			
woman	0.2	0.1		
number			0.1	
price			0.3	0.1

Parameters of trained model (randomly initialized, 10 re-estimation steps):

class 0	0.5	woman 0.6	man 0.4	smile 0.8	laugh 0.2
class 1	0.5	price 0.8	number 0.2	increase 0.8	fall 0.2

Toy Example: Empirical versus Model Probabilties

$\tilde{p}(y_1, y_2)$	smile	laugh	increase	fall	$\tilde{p}(y_1)$
man	0.2				0.2
woman	0.2	0.1			0.3
number			0.1		0.1
price			0.3	0.1	0.4
$\tilde{p}(y_2)$	0.4	0.1	0.4	0.1	

$p(y_1, y_2)$	smile	laugh	increase	fall	$p(y_1)$
man	0.16	0.04			0.2
woman	0.24	0.06			0.3
number			0.08	0.02	0.1
price			0.32	0.08	0.4
$p(y_2)$	0.4	0.1	0.4	0.1	

Comparing the empirical distribution (left) with the model distribution (right), two events attract attention:

- **smoothing takes place**: *(man, laugh)* and *(number,fall)* but not for e.g. *(man, increase)* or *(number,smile)*,
- **invariant marginal distributions** occure, i.e. $\tilde{p}(y_i) = p(y_1)$ $(i = 1, 2)$. **EM-based multivariate clustering preserves** the marginal distributions and thus some **characteristical properties of the given corpus**.

Invariant Marginal Distributions

Theorem: EM-based multivariate clustering leads to invariant marginals.
Proof: Using EM-based re-estimated probabilities $(i = 1 \ldots d)$:

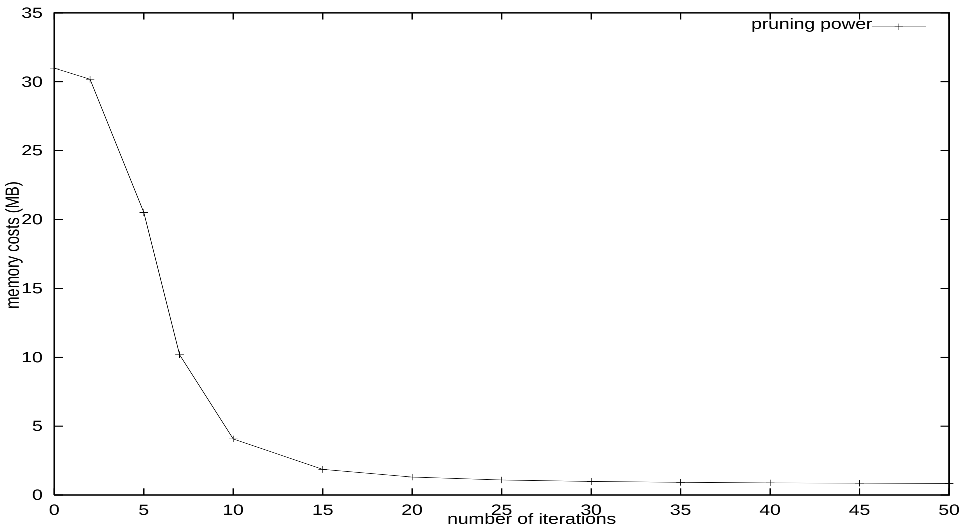
$$
\begin{aligned}
\hat{p}(y_i) &= \sum_{c \in C} \hat{p}(c) \cdot \hat{p}(y_i|c) \\
&= \sum_{c \in C} |f|^{-1} \cdot |f_c| \cdot |f_c|^{-1} \cdot f_c(y_i) \\
&= \sum_{c \in C} |f|^{-1} \sum_{y \in \mathcal{H}(y_i)} f_c(y) \\
&= \sum_{y \in \mathcal{H}(y_i)} |f|^{-1} \cdot f(y) \sum_{c \in C} p(c|y) \\
&= \sum_{y \in \mathcal{H}(y_i)} \tilde{p}(y) \qquad = \tilde{p}(y_i) \qquad \textbf{q.e.d.}
\end{aligned}
$$

Pruning Experiments

Pruning: after each re-estimation step, we deleted all data types with **infrequent class-based frequencies** to decrease the required storage space for EM-based multivariate clustering models:

$$f_c(y) < \epsilon \cdot |C|^{-1} \qquad (\epsilon := 0.001 \text{ in the implementation}).$$

Results: for a model with 35 classes of 27 000 verb-object pairs (WSJ)

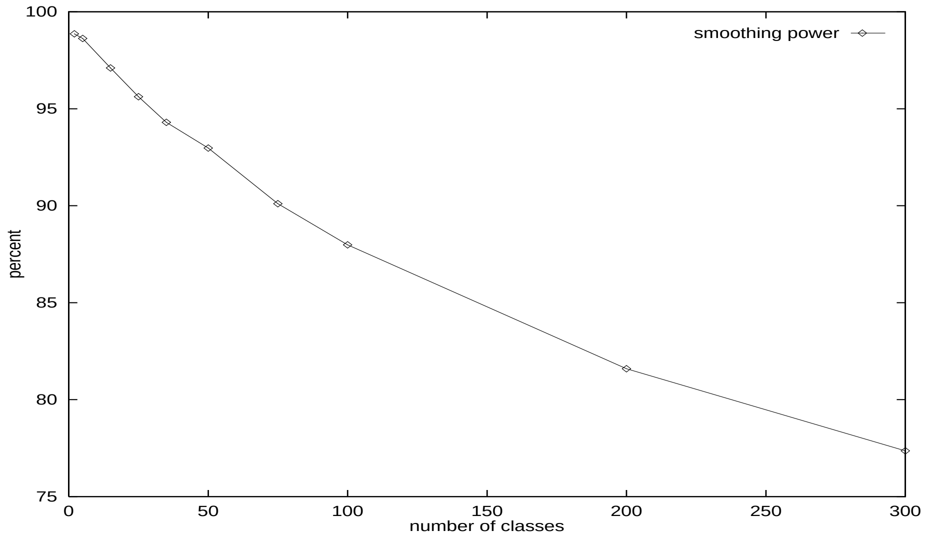


Smoothing Experiments

Smoothing power: percentage of data types having a positive probability.
We expect for **EM-based clustering models**:

- One-class models are **perfect smoothers** since $p(y) = \prod_i \tilde{p}(y_i)$
- In general, **good smoothing behaviour** since $p(y) = \sum_c p(c) \cdot p(y|c)$
- **Pruning decreases the smoothing power** since it delets data

Results: for models trained on 1 280 000 verb-noun pairs (BNC) with an initial smoothing power of 0.1%



Well-Tried Performance

Well-Tried Performance: EM-Based Multivariate Clustering (EMMC) was evaluated in **real-world applications** on a large number of randomly selected examples of a **real-world corpus**. Here is an overview:

NLP application	**Improvement of performance using EMMC**
Induction of semantically annotated lexicons	∞
Identification of collocations/idioms	0%
Grapheme-to-phoneme conversion	3% (30%)
Machine translation	7 %
Stochastic lexicalized parsing	13-16 %

Unsupervised Induction of Semantically annotated Lexicons

Reference: Rooth, Riezler, Prescher, Carroll, Beil (1999). Inducing a semantically annotated lexicon via EM-based clustering. In *Proceedings of ACL-99*.

Application/Method: Unsupervised induction of slot annotations for subcategorization frames, where annotations consist of the classes of an **EM-based 2-dimensional clustering model** trained on more than 1 000 000 verb-noun pairs gathered by stochastically parsing the British National Corpus (a free text corpus of about 1 00 000 000 words).

Gain (∞): We attribute an infinite gain of performance to EM-based clustering models, since **unsupervised induction** of semantically annotated lexicons completely relies on EM-based clustering models which can be automatically built using a statistical parser and free text.

Identification of Collocations/Idioms

Reference: Prescher, Heid (1999). Probabilistisches Clustering zur Identifikation von Verb-Nomen-Kollokationen. Presented at *DGfS-00*.

Application/Method: Identification of verb-noun collocations/idioms based on a simple comparison of the **empirical distribution** with the distribution of an **EM-based 2-dimensional clustering model** trained on about 500 000 verb-noun pairs gathered by parsing a German corpus of about 4 000 000 words.

Gain (0%): Recent evaluation on 400 items shows that this method yields as good results as current state-of-the-art methods.

Grapheme-to-Phoneme Conversion

Reference: Müller, Möbius, Prescher (2000). Inducing probabilistic syllable classes using multivariate clustering. In *Proceedings of ACL-00*.

Application/Method: Grapheme-to-phoneme conversion (G2P) was performed (i) using a context-free grammar to produce all phonemic correspondences of a given grapheme string, (ii) applying an **EM-based 3- or 5-dimensional clustering model** trained on syllables, and (iii) predicting pronunciation by choosing the most probable analysis.

Gain (3% resp. 30%): On about 2 000 test items, the G2P system using EM-based 5-dimensional clustering models achieves an increase over the baseline system using the **empirical distribution**. Moreover, it achieves an increase of about 30%, if compared to the **standard probability model** for context-free grammars.

Machine Translation

Reference: Prescher, Riezler, Rooth (2000). Using a probabilistic class-based lexicon for lexical ambiguity resolution. In *Proceedings of COLING-00*.

Application/Method: In machine translation, one problem to be solved is to find a correct translation of a source word using only minimal contextual information. The method is, to use the **class-based frequencies** of an EM-based 2-dimensional clustering model of verb-noun data to disambiguate alternative target nouns given their governing verb.

Gain (7%): Evaluation on a corpus with about 800 bilingual sentences yields an increase of 7% precision over the baseline system using the **empirical verb-noun distribution** as disambiguator.

Stochastic Lexicalized Parsing

Reference: Riezler, Prescher, Kuhn, Johnson (2000). Lexicalized stochastic modeling of constrait-based grammars using log-linear measures and EM training. In *Proceedings of ACL-00*.

Application/Method: Stochastic modeling of an unification grammar was performed using free text and EM training for log-linear models. With an EM-based 2-dimensional clustering model of verb-noun data, all parses of a sentence were pre-disambiguated according to **class-based frequencies** of the lexical heads in certain grammatical relations.

Gain (13%-16%): Evaluation on a corpus of about 500 items shows that incorporation of the pre-disambiguation routine into the log-linear model improves precision of the stochastic model by about 13%. Moreover, it improves the model by 16%, if simple treebank training is used instead of EM-training.

Discussion

- In contrast to similiar approaches (Rooth 1995; Pereira, Tishby, Lee 1995; Hofmann, Puzicha 1998), EM-based multivariate clustering is formalized clearly as pure EM algorithm, which plays a major role in the derivation of the invariance property.

- The approaches of (Schütze 1992, Yarowsky 1995) to word sense disambiguation use large amounts of information in terms of large context windows, which is easily obtainable in IR applications, but often unavailable in parsing or translation.

- Previous work on smoothing has shown that most of the developed methods are highly effective. However, EM-based multivariate clustering models utilizes the hidden structural information of given data and this is a feature lying beyond the capability of a pure smoother.

Conclusion

- **Simplified re-estimation formulas** on the basis of **class-based frequencies** lead to detection of a **new invariance property** of EM-based multivariate clustering,

- a **new pruning technique** relying on **class-based frequencies** makes EM-based multivariate clustering feasible and results in models of only twice the size of the original data,

- we motivated theoretically and showed experimentally that EM-based multivariate clustering models are **good smoother**,

- **EM-based multivariate clustering models** improve several NLP applications, where other clustering methods are not applicable or competing smoothing methods are useles.