

Semantic based DCM models for text classification ¹

Paola Cerchiello² and Elvio Concetto Bonafede
Dipartimento di Statistica ed Economia Applicate “Libero Lenti”,
Università di Pavia, e-mail: paola.cerchiello@unipv.it

Abstract:

This contribution deals with the problem of documents classification. The proposed approach is probabilistic and it is based on a mixture of a Dirichlet and Multinomial distribution. Our aim is to build a classifier able, not only to take into account the words frequency, but also the latent topics contained within available corpora. This new model, called *sb-DCM*, allows us to insert directly the number of topics (known or unknown) that compound the document, without losing the “burstiness” phenomenon and the classification performance. The distribution is implemented and tested according to two different contexts: on one hand, the number of latent topics is defined by experts in advance and, on the other hand, such number is unknown.

Keywords: Mixture models, Text analysis, Latent topics

1. Introduction

In the Dirichlet Compound Multinomial (or DCM) distribution (see Minka 03, Madsen 05) the Dirichlet random variable is devoted to model the Multinomial word parameters θ (the whole set of words as bag-of-bags) and the Multinomial variable models the word count vectors ($\bar{x}$ as bag-of-words) for each document. The distribution function of the DCM mixture model is:

$$p(\bar{x}|\alpha) = \int_{\theta} p(\bar{x}|\theta)p(\theta|\alpha)d\theta, \quad (1)$$

where the $p(\bar{x}|\theta)$ is the Multinomial distribution and $p(\theta|\alpha)$ is the Dirichlet distribution. From another point of view, each Multinomial is linked to specific sub-topics and makes, for a specific document, the emission of some words more likely than others. Instead the Dirichlet represents a general topic that compounds the set of documents and thus the DCM could be also described as “bag-of-scaled-documents”.

The added value of the DCM approach consists in the ability to handle the “burstiness” of a rare word without introducing heuristics. In fact, if a rare word appears once along a text, it is much more likely to appear again.

2. A Semantic-Based DCM

In the DCM we have a coefficient (α_w) for each word compounding the vocabulary of the set of documents which is called “corpus”. The DCM model can be seen as a “bag-of-scaled-documents” where the Dirichlet takes into account a general topic and the Multinomial some specific sub-topics.

¹This work has been supported by MUSING 2006 contract number 027097, 2006-2010.

Our aim in this contribution is to build a framework that allows us to insert specifically the topics (known or latent) that compound the document, without losing the “burstiness” phenomenon and the classification performance. Thus we introduce a method to link the α coefficients to the hypothetic topics, indicated with $\beta = \{\beta_j\}$, by means of a function $\alpha = F(\beta)$ which must be positive in β since the Dirichlet coefficients are positive. Note that usually $\dim(\beta) \ll \dim(\alpha)$ and, therefore, our proposed approach is parsimonious. See Bonafede Cerchiello Giudici '09 for more details.

Substituting the new function into the integral in (1), the new model is:

$$p(\bar{x}|\beta) = \int_{\theta} p(\bar{x}|\theta) p(\theta|F(\beta)) d\theta. \quad (2)$$

By substituting this linear combination into integral (2), we obtain:

$$p(\bar{x}|\beta) = \frac{n!}{\prod_w^M x_w} \frac{\Gamma(\sum_w^M \sum_t^R d_{wt}^* \beta_t)}{\Gamma(\sum_w^M (x_w + \sum_t^R d_{wt}^* \beta_t))} \prod_w^M \frac{\Gamma(x_w + \sum_t^R d_{wt}^* \beta_t)}{\Gamma(\sum_t^R d_{wt}^* \beta_t)}; \quad (3)$$

This model is a modified version of the DCM, henceforth semantic-based DCM (sb_DCM). An important aspect of the proposed approach is represented by the number R of topics to be inserted into the semantic-based DCM. R can be:

1. a priori known (i.e. fixed by field experts);
2. unknown, so to be estimated on the basis of the available data.

We treat both the cases, reporting empirical results produced on the basis of real data.

3. Semantic-Based DCM with R unknown

Since it is not always possible to known in advance the number of latent topics present in a corpora, it becomes very useful to arrange a statistical methodology for discovering efficiently and consistently a suitable number of topics.

In our context we tackle the problem as follows: in order to generate the splitted coefficients contained within the matrix **D** we have used a segmentation procedure to group the words. The idea is to create groups of words sharing common characteristics that can be considered latent topics. Such objective can be accomplished by applying a grouping analysis on the correlation matrix calculated on the complete set of words count. Later on, once completed the analysis by choosing the best number of groups, the cluster distance matrix is used to set the membership percentage (d_{wt}) of each word to each latent topic. Two kinds of matrixes have been used in the grouping procedure. One matrix contains the correlations among words (C) in the vocabulary and another one (G) is constructed by calculating the Kruskal-Wallis index on the count matrix among words. The training dataset contains 2051 documents with a vocabulary of 4096 words for both approaches. The evaluation dataset (again the same for both models) contains 1686 documents which are distributed over 46 classes. In Tables 1 and 2 we report the results from the two tests obtained respectively by matrixes **C** and **G** and by varying the number of groups in the cluster. In the tables we indicate with LL_{in} the log-likelihood before the parameters updating and with LL_{out} after the iteration procedure which is stopped when the error $\epsilon 10^{-10}$ is reached. The same with AIC_{cin} and AIC_{cout} that is the corrected Akaike Information Criterion (AICc) before and after the uploading.

Table 1:*Classification results by varying cluster numbers and using matrix (C)*

Classifier	Measures	sbDCM_5	sbDCM_11	sbDCM_17	sbDCM_23	sbDCM_46	DCM
	LL_{in}	-282226	-265250	-252197	-247125	-242991	-222385
	LL_{out}	-205412	-205614	-205601	-205597	-205602	-205286
	$AICc_{in}$	564462	530522	504228	494296	486074	454264
	$AICc_{out}$	410834	411250	411236	411240	411296	420066
NORMAL	$Ind1$	68.13%	68.13%	67.95%	68.19%	68.13%	67.66%
COMP.	$Ind1$	68.19%	68.31%	68.25%	68.37%	68.25%	68.78%
MIXED	$Ind1$	68.07%	68.36%	68.43%	68.37%	68.31%	68.07%

Table 2:*Classification results by varying cluster numbers and using matrix (G)*

Classifier	Measures	sbDCM_5	sbDCM_11	sbDCM_17	sbDCM_23	sbDCM_46	DCM
	LL_{in}	-291257	-283294	-270360	-266453	-258061	-222385
	LL_{out}	-205912	-204647	-204600	-204604	-204362	-205286
	$AICc_{in}$	582524	566610	540754	532952	516214	454264
	$AICc_{out}$	411834	409316	409234	409254	408816	420066
NORMAL	$Ind1$	67.83%	67.71%	67.47%	67.42%	67.65%	67.66%
COMP.	$Ind1$	67.95%	68.66%	68.55%	68.72%	68.60%	68.78%
MIXED	$Ind1$	68.07%	68.13%	67.83%	67.71%	67.89%	68.07%

As we can see in the two Tables 1 and 2, the percentages of correct classification ($Ind1$) are very close to the original ones with a parameter for each word (4096 parameters). Of course they depend on the type of classifier employed during the classification step. Considering both sbDCM and DCM, the differences produced by varying the number of groups are small. Moreover the AICc is always better in the new approach then considering each word as a parameter (DCM model).

4. Semantic-Based DCM with R known in advance

A different approach needs to be assessed when the number of available topic R is known in advance. In fact a text corpora could be enriched by several descriptions of treated topics according to the experience of the field experts.

In more details, the analysis could be provided with a priori knowledge based on ontological schemas that describe the relations among concepts with regards to the general class argument of the corpora. For example, if a text set deals with the operational risk management, we know that some specific sub-topics such as the general schema of possible operational problems will be almost surely treated in the texts. Thereby, the ontology structure provides the set of possible relations among the concepts to which can be associated, by a field expert(s), a certain number of key words.

In this context we employ the sb_{DCM} to insert directly the R topics. We have used a set of 870 articles about Alitalia, covering a period of one year (sept 07-sept08). The 80% of the articles are used to train the model and the 20% to validate the process. The objective is to

classify the articles on the basis of the reputation ontology in order to understand the argument treated in the articles. The ontology classes used for the classification are: *Identity, Imagination, Integrity, Quality, Reliability, Reputation, Social Responsibility, Technical Innovation, Value For Money*. Those classes defines the concept of reputation of an enterprise. To link the ontology classes to the textual analysis we use a set of key words for each class of the reputation schema. Since the articles are in Italian, the key words are in Italian as well. For example the concept of Reliability which definition is the ability of a system to perform/maintain its functions in routine and also in different hostile or/and unexpected circumstances. It involves customer satisfaction and customer fidelization, is characterized by the following set of key words: *affidabilità, fiducia, consumatori, risorsa, organizzazione, commerciale, dinamicità, valore, mercato*. On the basis of these key words, we perform a grouping analysis considering the 9 classes. From the clustering, we derive the matrix **D** (see Bonafede et al (2009)).

First empirical results show good performance in terms of correct classification rate. We have that over 171 articles we classify correctly the 68% of them. In figure 1 there is an example of output of classification with the coefficients α calculated via sbDCM and the Naive Bayes Classifier.

CLASS	DATA	TESTO
5	03/10/2007	dariole LEPIDO milano credo vendita alitalia realizzata entro anno alejandro bianchi ottimista ieri margine presentazione finanziaria settore trasporti risposto cronisti chiedevano lumi finale partita riguarda compagnia bandiera breve prossime settimane giorni aggiunto ministro trasporti presidente maurizio prao dovrebbe portare romano PRODI esito istruttoria vendita alitalia ministro bianchi parlato necessità incontro aziende sindacati richieste legittime ato dovuto perché tavolo avviato anno alimentato intanto air FRANCE KLM considerata prima candidata rilevare alitalia smorza entusiasmi sapere qualsiasi operazione gruppo alitalia snobismo vicepresidente compagnia van WUJK dovrà essere una alleanza con VALORE salvataggio politico resta vedere una alleanza realmente vantaggiosa visto parlano compagnia pesantemente indebitata necessita ricorganizzazione forse possa piano situazione potrebbe cambiare van WUJK inoltre smentito corso negoziati società private equity una offerta congiunta alitalia sottolineato una alleanza air FRANCE alitalia IBERIA verrebbe bloccata antitrust europeo sul fronte nord continua offensiva presidente lombardia roberto formigoni oggi presiederà roma riunione parlamentari lombardi lunedì prossimo convocato conferenza autonomie confronto province comuni governo inadempiente costretti alzare tono voce detto governatore lombardia mentre passati dodici giorni segreteria tecnica stata convocata governo può pensare tema lasciamo prendere stanchezza parlare voce chiara sempre ieri sottoscritto documento congiunto regione lombardia parti sociali economiche tema MALPENSA preservare caratteristiche hub sole varesino rafforzare collaborazione principali aeroporti nord italia ottica costruire network competitivo tra firmatari promotori patto SVILUPPO regione lombardia sindacati confindustria associazioni commerciali artigiani regolatori PMI mondo cooperativo

CLASSE 8: TECHNICALINNOVATION CLASSE 4: QUALITY; CLASSE 5: RELIABILITY

Figure 1: Example of classification via sbDCM and a Naive Bayes Classifier. The key words belong to three different classes: technical innovation; quality; reliability.

References

- Bonafede E.C., Cerchiello P. and Giudici P. (2009) Dirichlet compound multinomials for text modelling, *Technical Report*.
- Madsen R., Kauchak D. and Elkan C. (2005) Modeling word burstiness using the dirichlet distribution, *Proceeding of the 22st International Conference on Machine Learning*.
- Minka T. (1998) Expectation-maximization as lower bound maximization, *Technical Report*.
- Minka T. (2003) Estimating a dirichlet distribution, *Technical Report*.