Improving Classification of Documents by Semi-supervised Clustering in a Semantic Space

Jasminka Dobša and Henk A. L. Kiers

Abstract In the paper we propose a method for representation of documents in a semantic lower-dimensional space based on the modified Reduced k-means method which penalizes clusterings that are distant from classification of training documents given by experts. Reduced k-means (RKM) enables simultaneously clustering of documents and extraction of factors. By projection of documents represented in the vector space model on extracted factors, documents are clustered in the semantic space in a semi-supervised way (using penalization) because clustering is guided by classification given by experts, which enables improvement of classification performance of test documents.

Classification performance is tested for classification by logistic regression and support vector machines (SVMs) for classes of Reuters-21578 data set. It is shown that representation of documents by the RKM method with penalization improves the average precision of classification by SVMs for the 25 largest classes of Reuters collection for about 5,5% with the same level of average recall in comparison to the basic representation in the vector space model. In the case of classification by logistic regression, representation by the RKM with penalization improves average recall for about 1% in comparison to the basic representation.

Keywords: classification of textual documents, LSA, reduced k-means

Jasminka Dobša (✉)
Faculty of Organization and Informatics, University of Zagreb, Pavlinska 2, 40000 Varaždin, Croatia, e-mail: `jasminka.dobsa@foi.hr`

Henk A. L. Kiers
Department of Psychology, University of Groningen, Grote Kruisstraat 2/1, 9712 TS Groningen, The Netherlands, e-mail: `h.a.l.kiers@rug.nl`

P. Brito et al. (eds.), *Classification and Data Science in the Digital Age*,
Studies in Classification, Data Analysis, and Knowledge Organization,
https://doi.org/10.1007/978-3-031-09034-9_14

1 Introduction

There are two main families of methods that deal with representation of documents and words that index them: global matrix factorization methods such as Latent Semantic Analysis (LSA) [2] and local context window methods such as the continuous bag of words (CBOW) model and the continuous skip-gram model [8]. The latter use neural networks for learning of representations of words and are intensively explored lately in the scientific community since the development of fast processors has enabled processing of huge amounts of data which resulted in improvements in performance of wide spectra of text mining and natural language tasks. However, representation of words solely by context window methods has a drawback due to the neglect of information about global corpus statistics [9].

In this paper we propose a method for representation of documents by application of a penalized version of the RKM method [4] on a term-document matrix. The corpus of textual documents is represented by a sparse term-document matrix in which entry (i, j) is equal to the weight of the i-th index term for the j-th document. Weights of terms are given by the TfIdf weighting which utilizes local information about the frequency of the i-th term in the j-th document and global information about usage of the i-th term in the entire collection. A benchmark method that utilizes global matrix factorization on term-document matrices is LSA [2] which uses truncated singular value decomposition (SVD) for representation of terms and documents in lower-dimensional semantic space. SVD does not capture the clustering structure of data which motivates application of the RKM.

The rest of the paper is organized as follows: the second section describes related work on representation of documents and words and methods of dimensionality reduction related to RKM. The third section describes the modified RKM method with penalization, while the fourth section describes an experiment on Reuters-21578 data set. In the last section conclusions and directions for further work are given.

2 Related Work

2.1 Representation by Matrix Factorization Methods

A benchmark method among methods that utilize matrix factorization for representation of textual documents is the method of LSA introduced in 1994 [2]. By LSA a sparse term-document matrix is transformed via SVD into a dense matrix of the same term-document type with representations of words (index terms) and documents in a lower-dimensional space. The idea is to map similar documents, or those that describe the same topics, closer to each other regardless of the terms that are used in them. A very efficient application of LSA is in cross-lingual information retrieval where relevant documents for a query in one language are retrieved from a set of documents in another language [7]. According to our knowledge application of methods that simultaneously cluster objects and extract factors in the field of text mining is very limited. In [6] a method is proposed for cross-lingual information retrieval based on the RKM method.

2.2 Neural Network Word Embeddings

Another approach is to learn representations of words, or so called embeddings, by using local context windows. In 2003 Bengio and coauthors [1] proposed a neural probabilistic language model that uses simple neural network architecture to learn distributed representations for each word as well as probability functions for word sequences, expressed in terms of these representations. Mikolov and coautors [8] proposed in 2013 two models based on single-layer neural network architectures: the skip gram-model that predicts context words given the current word and the continuous bag of words model which predicts current words based on the context. In 2014 the GloVe model [9] was proposed, based on the critique that neural network models suffer from the disadvantage that they do not utilize co-occurrence statistics of the entire corpus, but scan only context windows of words ignoring vast amounts of repetition in the data. That model exploits the advantages of global matrix factorization methods by utilization of term-term co-occurrence matrices and local context window methods.

Word embedding can be classified as static such as word2vec [8] and GloVe [9], and contextual, such as ELMo [10] and BERT [5]. Contextual representation is introduced in [10] in order to model characteristics of word use (syntax and semantics) on one side and variation in word representation due to the context in which words are appearing.

2.3 Methods for Simultaneous Clustering and Factor Extraction

A standard procedure for clustering of objects in a lower-dimensional space is tandem analysis which includes projection of data by principal components and clustering of data in a lower-dimensional space. Such an approach was criticized in [3] and [4] since principal components may extract dimensions which do not necessarily significantly contribute to the identification of a clustering structure in the data. As a response, De Soete and Carroll proposed the method of RKM [4] which simultaneously clusters data and extracts the factors of variables by reconstructing the original data with only centroids of clusters in a lower-dimensional space. The algorithm of Factorial k-means (FKM) proposed by Vichi and Kiers [13] has the same aim of simultaneous reduction of objects and variables and it reconstructs the data in a lower-dimensional space by its centroids in the same space. The application of the latter method is limited in text mining since the method is limited to cases in which the number of variables is less than the number of cases. In [11] the RKM and FKM methods are compared using simulations and theoretically in order to identify cases for their application. Timmerman and associates also propose method of Subspace k-means [12] which gives an insight into cluster characteristics in terms of relative positions of clusters given by centroids and the shape of the clusters given by within cluster residuals.

3 Reduced k-Means with Penalization

Let $\mathbf{X}$ be $m \times n$ term-document matrix. We use the following notation:

- $\mathbf{A}$ is an $m \times k$ columnwise orthonormal matrix of extracted factors;
- $\mathbf{M}$ is an $n \times c$ membership matrix, where c is a predefined number of clusters; $m_{ic} = 1$ if object (document) i belongs to cluster c and 0 otherwise;
- $\mathbf{Y}$ is a $c \times k$ matrix which gives centroids of clusters in the lower-dimensional space.

By definition, we suppose that every document in the collection belongs to exactly one cluster. The RKM method minimizes the loss function

$$\mathbf{F(M,A)} = ||\mathbf{X} - \mathbf{A}\mathbf{Y}^T\mathbf{M}^T||^2 \tag{1}$$

in the least squares sense. The dimension of the lower-dimensional space must be less or equal to the number of clusters. Modified RKM with penalization minimizes the loss function

$$\mathbf{F(M,A)} = ||\mathbf{X} - \mathbf{A}\mathbf{Y}^T\mathbf{M}^T||^2 + \lambda||\mathbf{M} - \mathbf{G}||^2 \tag{2}$$

where $\mathbf{G}$ is $n \times c$ membership matrix based on expert judgements. If c is number of classes then $g_{ic} = 0$ if object (document) i belongs to class c, and 0 otherwise. By the second summand in the loss function we penalize clusterings that are distant from the classes by expert judgements using parameter λ that regularizes the importance of that penalization. We use the alternating least squares (ALS) algorithm analogous to the one in [4] which alternates between corrections of the loading matrix $\mathbf{A}$ in one step and of the membership matrix $\mathbf{M}$ in another. As each of the steps in the ALS algorithm improves the loss function, the algorithm converges to at least a local minimum. By starting the procedure from a large number of random initial estimates and choosing the best solution, the chances of obtaining the global minimum are increased.

4 Experiment

4.1 Design of Experiment

Experiments are conducted for classification on the Reuters-21578 data set, specifically using the ModApte Split which assigns Reuters reports from April 7, 1987 and before to the training set, and after, until end of 1987, to the test set. It consists of 9603 training and 3299 test documents. The collection has 90 classes which contain at least one training and test document. Documents are represented by a bag of words representation. A list of index terms is formed based on terms that appear in at least four documents of the collection, which resulted in a list of 9867 index terms.

Classification is conducted by logistic regression (LR) and SVM algorithm. The basic model is the bag of words representation (full representation), while representations in the lower-dimensional space are obtained by SVD (Latent Sematic Analysis), RKM and RKM with penalization ($\lambda = 0.1, 0.2, 0.4, 0.6$). For RKM and RKM with penalization representations are obtained by applying matrix factorization on the term-document matrix of the training documents, and by projection of test documents on factors given by matrix A in the factorization. RKM is computed for 90 clusters (which corresponds to the number of classes in the collection) using as dimension of the lower-dimensional space $k = 85$, and truncated SVD is computed for $k = 85$ as well. The RKM and RKM with penalization algorithms are run 10 times (with different starting estimates), and the representation and factorization with the minimal loss function is chosen. The optimal cost parameter for LR and SVM is chosen by grid search technique from the set of values 0.1, 0.5, 1, 10, 100 and 1000. For the classification methods, the LiblineaR library in R is used, while RKM and RKM with penalization algorithm are implemented in Matlab.

4.2 Results

Results are given in terms of precision, recall, and F_1 measure of the classification. Recall is proportion of correctly classified samples among all positive samples (i.e., samples actually belonging to the class, according to the expert), while precision is proportion of correctly classified samples among all samples classified as positive by the model. In the Figures 1 and 2, are shown results of average F_1 measures of classification for 5 classes sorted in descending order by their size, i.e. number of train documents (which is 2877 to 389 for classes 1-5, 369 to 181 for classes 6-10, 140 to 111 for classes 11-15, 101 to 75 for classes 16-20, 75 to 55 for classes 21-25, 50 to 41 for classes 26-30, 40 to 37 for classes 31-35, 35 to 24 for classes 36-40, 23 to 19 for classes 41-45, 18 to 16 for classes 46-50, 16 to 13 for classes 51-55, and 13-10 for classes 56-60). Figure 1 shows the results for classification by LR, while Figure 2 for classification by SVM. Only the 60 largest classes are observed since smaller classes (less than 10 training documents) are not interesting for the

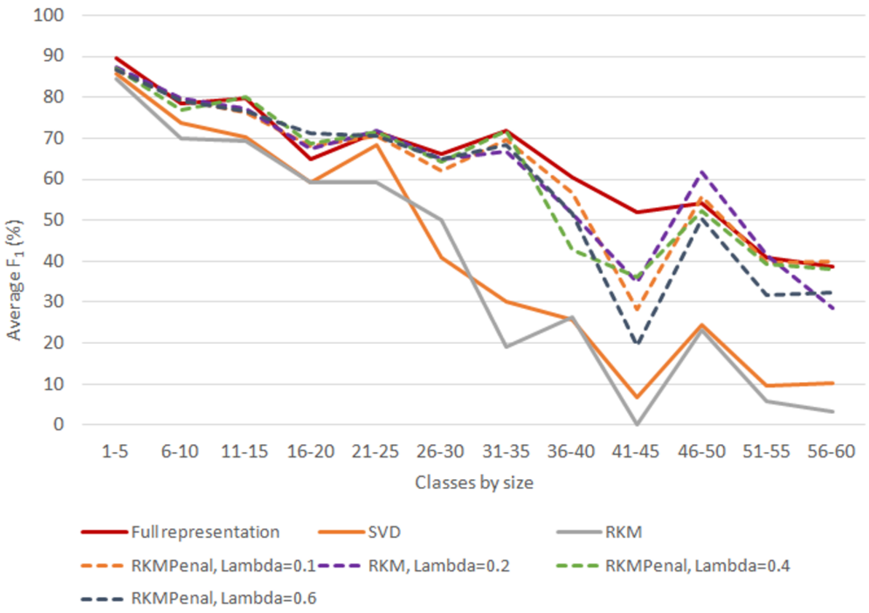

Fig. 1 Average F_1 measure of classification by LR for 5 classes sorted by their size.

research, because for those classes recall is low and it can be expected that full bag of words representation will result in better recognition since classes can possibly be recognized by key words, but not by transformed representations. It can be seen that F_1 measures are comparable for the full representation and various representations by RKM with penalization for both classification algorithms for the biggest 25 classes. For smaller classes results for representation by RKM with penalization are unstable, although for some classes they were better than the basic representation (in the case of LR). Classification for representations obtained by SVM and RKM without penalization resulted in lower F_1 measures for all class sizes.

In Table 1 are shown average precision, recall and F_1 measures for the 25 largest classes for both classification algorithms and all observed representations. In the case of classification by LR the average recall is improved for representation by RKM with penalization (for $\lambda = 0.4$) approximately 1% compared to basic full representation. For classification by SVM average precision is improved for representation by RKM with penalization (for $\lambda = 0.6$) for almost 6% and F_1 measure is improved for representation by RKM with penalization ($\lambda = 0.4$) for 2% in comparison to the basic full representation. The best results are obtained for classification by the SVM algorithm and representation with RKM with penalization with $\lambda = 0.2$ for which precision is improved for 5% with the similar level of recall as in the basic representation.

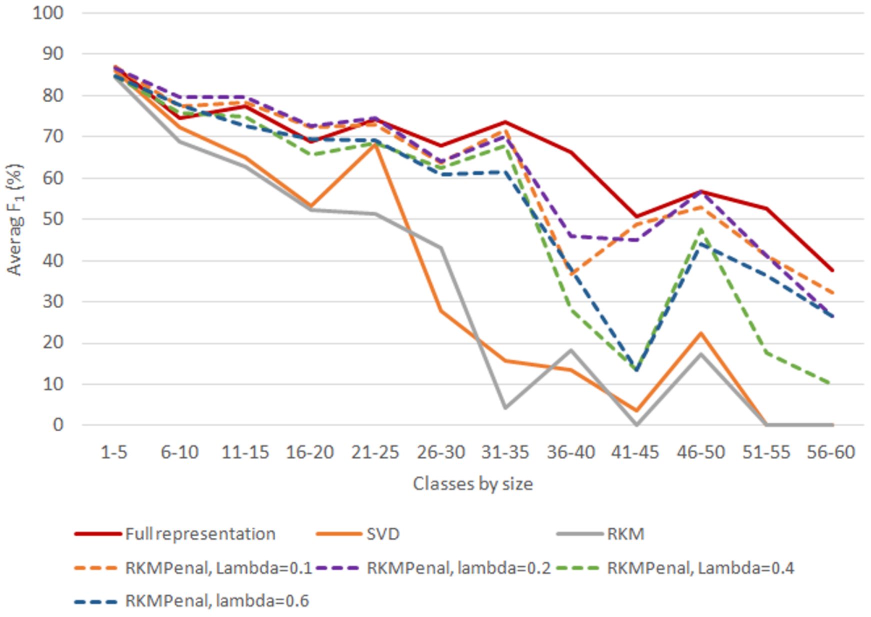

Fig. 2 Average F_1 measure of classification by SVM for 5 classes sorted by their size.

Table 1 Average precision, recall, and F_1 measure of classification for the 25 largest classes.

Class. algorithm	Logistic regression			SVM		
Representation	Precision	Recall	F_1	Precision	Recall	F_1
Full	**86.31**	70.24	76.84	82.76	71.72	76.47
SVD	82.80	64.84	71.42	85.24	61.61	68.99
RKM	80.80	61.10	68.44	82.93	55.66	63.83
RKMPenal, $\lambda = 0.1$	84.24	70.71	76.27	87.24	71.01	77.62
RKMPenal, $\lambda = 0.2$	84.68	71.23	76.72	87.78	**72.16**	**78.57**
RKMPenal, $\lambda = 0.4$	84.72	**71.38**	**76.88**	87.86	64.93	73.87
RKMPenal, $\lambda = 0.6$	85.89	70.40	76.80	**88.40**	66.11	74.75

5 Conclusions and Further Work

In this paper we propose a modification of the RKM method that simultaneously clusters documents and extracts factors on one side, and penalizes clusterings that are distant from the classification of the training documents given by experts on the other side. We show that such a modification enables representation of textual documents in a semantic lower-dimensional space that improves performance of classification. The method is tested for classes of Reuters-21758 data set and compared to the full bag of words representation and the method of LSA. It is also shown that the original RKM method without proposed modification does not have the same effect on classification performance; it has a similar effect as the LSA method.

The proposed representation method can improve precision and recall of classification for sufficiently large classes, i.e. those that have enough training documents to enable capturing of semantic relations and characteristics of classes. A more important effect can be observed in the improvement of precision.

In the future we plan to investigate hybrid models using representation of words by neural language models and application in different domains, such as classification of images.

References

1. Bengio, J., Ducharme, R., Vincet, P., Jauvin, C.: A Neural probabilistic language model. Journal of Machine Learning Research **3**, 1137-1155 (1997)
2. Deerwester, S., Dumas, S. T., Furnas, G.W., Landauer, T. K., Harshman, R. A.: Indexing by latent semantic analysis. Journal of the American Society for Information Science **41(6)**, 381-407 (1990)
3. De Sarbo, W. S., Jedidi, K., Cool, K., Schendel, D.: Simultaneous multidimensional unfolding and cluster analysis: an investigation of strategic groups. Marketing Letters, **2**, 129-146 (1990)
4. De Soete, G., Carroll, J. D.: K-means clustering in a low-dimensional Euclidean space. In: Diday, E., Lechevallier, Y., Schader, M., Bertrand, P., Burtschy, B. (eds.) New Approaches in Classification and Data Analysis. Studies in Classification, Data Analysis, and Knowledge Organization, pp. 212-219. Springer, Heidelberg (1994)
5. Devlin, J., Chang, M., Lee, K., Toutanova, K.: BERT: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of Annual Conference of the North American Chapter of the Association for Computation Linguistic, pp. 4171-4186, Association for Computational Linguistic (2019)
6. Dobša, J., Mladenić, D., Rupnik, J., Radošević, D., Magdalenić, I.: Cross-language information retrieval by Reduced k-means, International Journal of Computer Information Systems and Industrial Management Applications, **10**, 314-322 (2018)
7. Dumas, S., Letche, T., Littman, M., Landauer, T.: Automatic cross-language retrieval using latent semantic indexing. In: Proceedings of the AAAI spring symposium on cross-language text and speech retrieval, pp. 15-21. American Association for Artificial Intelligence (1997)
8. Mikolov, T., Chen, K., Corrado, G.S., Dean, J.: Efficient estimation of word representations in vector space (2013) Available via arXiv.org
`https://arxiv.org/abs/1301.3781.Cited21Jan2022`
9. Pennington, J., Socher, R., Manning, C. D.: GloVe: Global vectors for word representation. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 1532-1543, Association for Computational Linguistics, (2014)
10. Peters, M., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., Tettlemoyer, L.: Deep contextualized word representations. In Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 1:2227-2237 (2018)
11. Timmerman, M. E. Ceulemans, E., Kiers, H. A. L., Vichi, M: Factorial and Reduced k-means reconsidered. Computational Statistics & Data Analyisis, **54**, 1856-1871 (2010)
12. Timmerman, M. E., Ceulemans, E., De Rover, K., Van Leeuwen, K.: Subspace$k$-means clustering, Behavioural Research, **45**, 1011-1023 (2013)
13. Vichi, M., Kiers, H. A. L.: Factorial k-means analysis for two-way data, Computational Statistics & Data Analysis, **37**, 49-64 (2001)