






Distance Metrics in Open-Set Classification of Text Documents by Local Outlier Factor and Doc2Vec

Tomasz Walkowiak^(✉), Szymon Datko, and Henryk Maciejewski

Faculty of Electronics, Wrocław University of Science and Technology,
Wrocław, Poland
{tomasz.walkowiak,szymon.datko,henryk.maciejewski}@pwr.edu.pl

Abstract. In this paper, we investigate the influence of distance metrics on the results of open-set subject classification of text documents. We utilize the Local Outlier Factor (LOF) algorithm to extend a closed-set classifier (i.e. multilayer perceptron) with an additional class that identifies outliers. The analyzed text documents are represented by averaged word embeddings calculated using the fastText method on training data. Conducting the experiment on two different text corpora we show how the distance metric chosen for LOF (Euclidean or cosine) and a transformation of the feature space (vector representation of documents) both influence the open-set classification results. The general conclusion seems to be that the cosine distance outperforms the Euclidean distance in terms of performance of open-set classification of text documents.

Keywords: Text mining · Subject classification · Open-set classification · Word embedding · fastText · Local Outlier Factor · Cosine distance · Standarization

1 Introduction

The classification of texts, i.e. automatically assigning a text to one of predefined subject groups, becomes a tool useful in many areas (like digital libraries, newspaper repositories, categorization of scientific papers, questions, and answering systems, selection of tourist offers). However, practical usage of such a tool requires an extension of classical closed-set classifiers to open-set ones. Since a classical approach associates a new document to one of the trained classes, even if the document is actually not related to any of them, it can lead to very spectacular mistakes made by close-set text classification tools, for example assigning a random text to some class.

Therefore, we propose an extension to the standard closed-set classification schema. First, we build a standard classification model using the available training dataset. Next, we utilize the Local Outlier Factor (LOF) [1] algorithm to extend the result with an additional class that identifies outliers. LOF provides a measure of dissimilarity (outlierness factor), which proves useful for

high-dimensional data. Other approaches to open-set classification of text documents involve the utilization of statistical-based concepts, like inter-quartile range-based criteria, [13], or similarity estimation with simple threshold-based decision mechanism applied for the *a posteriori* probability [2, 3]. There are also approaches based on the usage of convolutional neural networks [11].

Moreover, we analyze how the distance metric chosen for LOF and a transformation of the feature space (vector representation of documents) influence the open-set classification results.

The paper is organized as follows. Section 2 describes the doc2vec representation of documents, the method of open-set classification and Local Outlier Factor algorithm used by this method. In Sect. 4 we discuss the used distance metrics and methods of feature space transformation. Next section presents the used corpora, experiments, and results of the comparative study.

2 Open Set Classification

2.1 Doc2vec

Several approaches to representing documents by feature vectors are available like classical bag-of-words and a number of its modifications. Recently, word embedding methods gained large popularity. They mostly act as a lookup table that maps each word into a continuous multidimensional vector space [8]. Word2Vec allows constructing a feature vector of an entire document (doc2vec) by simple average of word embeddings [5]. Within this paper, we have used a recent deep learning method – fastText [5].

The main idea behind fastText is to perform word embedding and classifier learning in parallel. FastText forms the linear model, since it consists of word embeddings, simple averaging and linear soft-max classifier. Therefore it is very effective to train and use.

2.2 Classification

Having doc2vec vectors we can train a typical (closed-set) classifier using standard machine learning algorithms based on the training set. Next, it can be used to assign any new document (described by doc2vec values) to one of classes occurring in the training set. In other words, the classifier splits the feature space into areas related to the trained classes. Hence, they associate a new document with one of the trained classes (winning class), even if the document is actually not related to any of the classes (subject categories) known to the classifier.

To overcome this problem we propose post-processing of typical classifier results. It includes calculation of dissimilarity between the feature vector and the winning class feature vectors (of all documents from the training set that belongs to the winning class). It could be done by measuring the outlierness factor [13]. If the factor is above given threshold we reject the decision made by the classifier and assign the document to the 'outlier' class.

2.3 Outlierness Factor

As an outlierness factor is required in the procedure described above, we propose to use the Local Outlier Factor (LOF) [1]. It is a measure based on a weighted Euclidean metric, aiming to find outliers by comparing vectors to their local neighborhoods. It works by calculating an average distance between a given point, its neighbors and their (neighbors) neighbors to determine the local density of points in the given point's surrounding.

The open-set classification procedure proposed in Sect. 2.2 requires thresholds for each class. We could set it up assuming that the training data sets are contaminated by outliers, i.e. they include a given proportion of vectors with LOF values larger than the threshold. This proportion is called contamination¹.

3 Analyzed Distance Metrics and Transformations of Feature Space

Original LOF [1] is based on Euclidean distance (often called L_2 norm). However, Beyer et al. [6] suggest that L_2 norm fails in high dimensions. Whereas, the *cosine distance* (mostly in the form of cosine similarity) is widely and successfully used in the analysis of word2vec data [7], as well as in doc2vec [10]. The cosine distance is not vulnerable to any scaling of the given vector's size and it is assumed that it acts much better in high dimensional space.

The *standardization*, i.e. removing the mean and scaling to the unit variance, is a widely used transformation of data in the machine learning. It allows matching the requirement of normal data distribution that is assumed in many classification algorithms. It is known in the statistics also as the z-score. It could be seen as moving and linear scaling of the input data. A mean and a standard deviation are calculated on training data and later on used during open-set classification.

The *normalization* scales the vector to the unit norm. In contrary to the standardization, it operates on an individual object and does not require any parameters estimated on other data. Normalization maps each data point onto unit n -sphere. It has some interesting properties. The cosine distance in original and normalized space are equal. Moreover, the cosine distance between vectors is equal to half of the square of Euclidean distance between them. Therefore, the results of algorithms based on the nearest neighbor (as LOF, for example) should be almost the same when someone uses cosine distance or Euclidean one on normalized vectors [12].

4 Evaluation

4.1 Data Sets

To evaluate the performance of the proposed open set classification method and analyze the influence of distance metrics and feature space transformation in

¹ <https://scikit-learn.org/0.19/modules/generated/sklearn.neighbors.LocalOutlierFactor.html>.

a real task the text corpora of different subject classes are needed. We have used two data sets: texts from English newsgroups (*20newsgroups*) and Polish Wikipedia (*Wiki*) articles.

The first corpus (*20newsgroups*) is a commonly used collection of nearly 20,000 forum posts² divided into 20 subject categories. The data were divided into training and testing sets. However, for the purpose of open set classification we need also an outlier data. For this purpose, we have selected following categories: *misc.forsale*, *talk.politics.misc*, *talk.politics.guns*, *talk.politics.mideast*, *talk.religion.misc* and *alt.atheism*, then we removed them from the training data.

The second corpus (*Wiki*) consists of ca. 10,000 Polish language Wikipedia articles [9], coming from 34 subject areas. In this case, the outlier data consists of randomly selected articles from Polish press news [14]. The number of outliers was equal to the size of a test partition.

4.2 Experiment Overview

The proposed method was evaluated on corpora described in Sect. 4.1. Firstly (for each corpus), the word2vec model was built using the fastText algorithm (Sect. 2.1) on the training set. Next, doc2vec feature vectors, as an average of word2vec values for each word in a document were calculated, forming the feature vector space. For closed-set classification, the multilayer perceptron (MLP) [4] with Broyden–Fletcher–Goldfarb–Shanno (BFGS) nonlinear optimization learning algorithm was used. The MLP model was built on the training set. Then, a constructed model was examined on both testing sets, labeling all documents to trained categories (closed-set classification). Later the Local Outlier Factor measure was used to verify if the assignment to categories was correct and to catch incorrect labels, marking mismatched data as outliers (open-set classification).

Finally, knowing all true labels from the original data-set, the evaluation of classification was performed. We measured a number of correct decisions from all assignments made to a specific class (precision) and a number of correct decisions from all assignments expected to a specific class (recall). Then, a harmonic mean of these values, called **f1-score** was calculated. The results, reported later, are given as the average of f1-scores for each class weighted by support (the number of instances in each class). It is important to notice that 50% of the testing data consists of outliers (for open-set data) so they have important influence on the final results.

4.3 Results

In a Table 1 we report the f1-score calculated for closed-set and open-set tasks for *20newsgroups* and *Wiki* corpora. The experiments were performed for 100 dimensional word embeddings and contamination parameter equal to 0.1. The second column is presented for a reference showing how the closed-set classifier performs in task dedicated for it (only the closed-set test data were used). Next

² <http://qwone.com/~jason/20Newsgroups/>.

columns show the results for open-set tasks (the closed-set test data and outliers are used). It can be noticed (second and third column) that introducing outlier data (50% of all documents) to a classical classification method (closed-set one) results in almost 2.5 time degradation of f1-score. However, the proposed by authors method (Sect. 2.2) is capable of improving the outcome (the fourth column) and achieve almost 63% or 54% of f1-score, depending on the outlier data set.

Next, we have analyzed the influence of feature vectors' transformation (none, standardization and normalization) and distance metric (Euclidean and cosine) on proposed method's performance (Table 2). It can be noticed that replacing a standard for LOF Euclidean distance with cosine one leads to the improvement of performance. Additional improvement, but only in case of *Wiki* corpus, is achieved when standardization is used. As it was mentioned in Sect. 3, the results for cosine distance with and without normalization are equal to the results obtained for Euclidean distance with normalization.

We have also tested other distance metrics available in SciPy³ package. The results are not shown since they were never better than for cosine one and many times even worse than for Euclidean metric.

Figure 1 shows the relation between contamination parameter of Local Outlier Factor (Sect. 2.3) and f1-score for *20newsgroups* and *Wiki* corpus respectively. We have shown (blue lines) the results for cosine and Euclidean distance, as well as cosine with standardization of vector space. For a reference, which

Table 1. Classification results: f1-score (word2vec dimension: 100, LOF contamination: 0.1)

Dataset	Closed-set	Open-set	
Method	Closed-set (MLP)	Open-set (MLP + LOF)	
20 newsgroups	0.7949	0.3073	0.6292
Wiki	0.8333	0.3119	0.5361

Table 2. Open set classification results: f1-score (word2vec dimension: 100, LOF contamination: 0.1)

Distance	Transformation	<i>20newsgroups</i>	<i>Wiki</i>
Euclidean	-	0.6292	0.5361
Euclidean	Standardization	0.6248	0.5358
Euclidean	Normalization	0.6512	0.7339
Cosine	-	0.6512	0.7335
Cosine	Standardization	0.6438	0.7902
Cosine	Normalization	0.6512	0.7335

³ <https://docs.scipy.org/doc/scipy/reference/spatial.distance.html>.

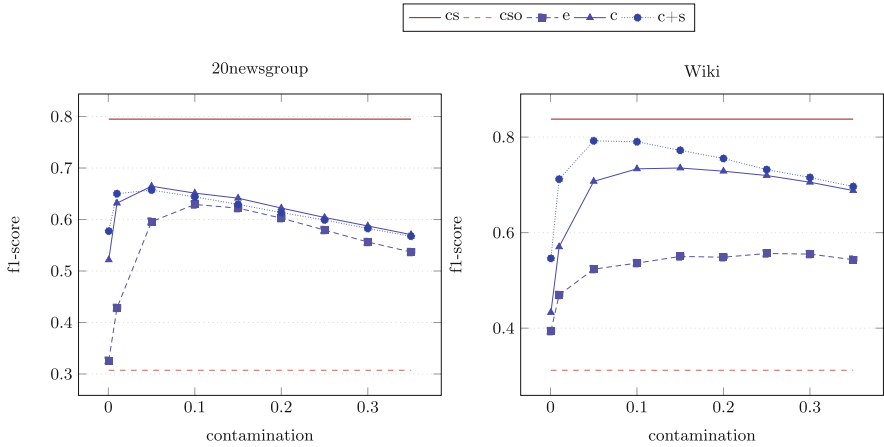


Fig. 1. f1-score for different values of contamination (cs - closed-set classifier without outliers, cso - closed-set classifier with open set data, e - open-set classifier with euclidean distance, c - open-set classifier with cosine distance, c+s - open-set classifier with cosine distance and standardization) (Color figure online)

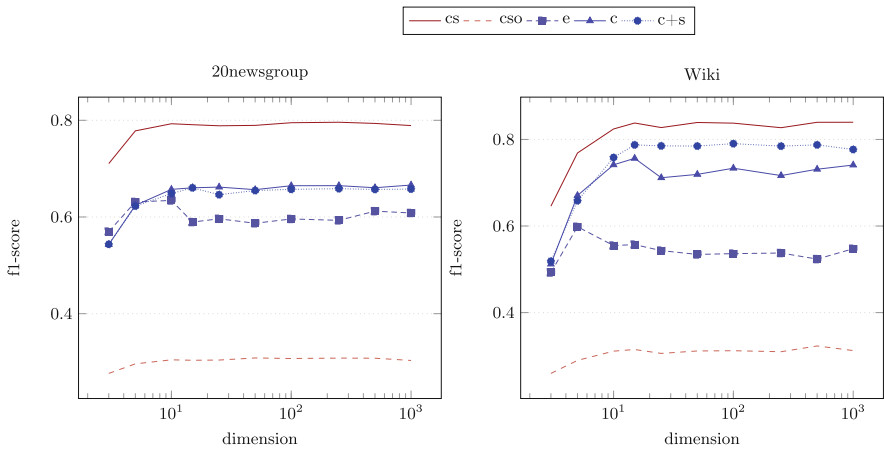


Fig. 2. f1-score in function of feature vector dimension (cs - closed-set classifier without outliers, cso - closed-set classifier with open set data, e - open-set classifier with euclidean distance, c - open-set classifier with cosine distance, c+s - open-set classifier with cosine distance and standardization)

could be interpreted as the bottom and top limit (red lines), we have also presented f1-scores for the closed-set task performed by the MLP on data with and without outliers. Obviously, these two do not depend on the contamination value (LOF is not involved). For the remaining three (open-set classification), when the values of contamination are rising, the f1-score is improving. Achieving its maximum at contamination value equal to 0.05 and 0.1 (only for *20newsgroups* and

pure Euclidean based LOF). After that, there is a decrease. Moreover, it could be noticed how cosine based LOF outperforms the Euclidean one, regardless of the data set and contamination value.

Next, we have analyzed the influence of word embedding dimensionality (range 3 – 1000) on the open-set classification performance. The results are presented in Fig. 2 for contamination parameter equal to 0.05 for *20newsgroups* and 0.1 for *Wiki*. The impact of doc2vec dimension for values larger than 10 appears not very significant. This may suggest that fastText algorithm is so effective in finding well-distinguishing features (word embeddings) that after some specific dimension there is only insignificant redundancy introduced.

Again, we can notice that cosine distance outperforms the Euclidean one. However, it is much more significant for *Wiki* corpus.

5 Conclusion

In this work, we showed how to extend the standard closed-set classifier (MLP was used during reported experiments) for the open-set classification of text documents described by doc2vec feature vectors. It is done by utilizing the Local Outlier Factor on document embeddings. In the experiment, we evaluated the proposed method on a collection of nearly 20,000 forum posts in English and Wikipedia articles in Polish (with 34 subject areas). The results show that the proposed extension is capable to work effectively in an open set environment.

Moreover, we researched various distance metrics and measured their performance in the task of open-set classification of text documents. Results show that using the cosine distance metric in LOF procedure we reach highest overall score on both examined datasets.

We have also studied the effect of two common transformations of feature vectors - standardization and normalization. In case of one of data sets, standardization allowed to boost results, whereas normalization gives the same results as a usage of cosine distance.

Acknowledgement. This work was sponsored by National Science Centre, Poland (grant 2016/21/B/ST6/02159).

References

1. Breunig, M.M., Kriegel, H.P., Ng, R.T., Sander, J.: LOF: Identifying density-based local outliers. *SIGMOD Rec.* **29**(2), 93–104 (2000). <https://doi.org/10.1145/335191.335388>
2. Doan, T., Kalita, J.: Overcoming the challenge for text classification in the open world. In: 2017 IEEE 7th Annual Computing and Communication Workshop and Conference (CCWC), pp. 1–7. IEEE (2017)
3. Fei, G., Liu, B.: Breaking the closed world assumption in text classification. In: Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 506–514 (2016)

4. Hastie, T., Tibshirani, R., Friedman, J.: *The Elements of Statistical Learning*. SSS. Springer, New York (2009). <https://doi.org/10.1007/978-0-387-84858-7>. Autres impressions : 2011 (corr.), 2013 (7e corr.)
5. Joulin, A., Grave, E., Bojanowski, P., Mikolov, T.: Bag of tricks for efficient text classification. In: *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pp. 427–431. Association for Computational Linguistics (2017)
6. Beyer, K., Goldstein, J., Ramakrishnan, R., Shaft, U.: When is “nearest neighbor” meaningful? In: *ICDT 1999 Proceedings of the 7th International Conference on Database Theory*, pp. 217–235 (1999)
7. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. *CoRR* abs/1301.3781 (2013). <http://arxiv.org/abs/1301.3781>
8. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. In: *Advances in Neural Information Processing Systems*, pp. 3111–3119 (2013)
9. Młynarczyk, K., Piasecki, M.: Wiki train - 34 categories, CLARIN-PL digital repository (2015). <http://hdl.handle.net/11321/222>
10. Pandey, N.: Density based clustering for cricket world cup tweets using cosine similarity and time parameter. In: *2015 Annual IEEE India Conference (INDICON)*, pp. 1–6 (2015). <https://doi.org/10.1109/INDICON.2015.7443520>
11. Prakhya, S., Venkataram, V., Kalita, J.: Open set text classification using convolutional neural networks. In: *Proceedings of the 14th International Conference on Natural Language Processing*, pp. 466–475. NLP Association of India, Kolkata (2017)
12. Qian, G., Sural, S., Gu, Y., Pramanik, S.: Similarity between Euclidean and cosine angle distance for nearest neighbor queries. In: *Proceedings of the 2004 ACM Symposium on Applied Computing, SAC 2004*, pp. 1232–1237. ACM, New York (2004). <https://doi.org/10.1145/967900.968151>
13. Walkowiak, T., Datko, S., Maciejewski, H.: Algorithm based on modified angle-based outlier factor for open-set classification of text documents. *Appl. Stochast. Models Bus. Ind.* **34**(5), 718–729 (2018)
14. Walkowiak, T., Malak, P.: Polish texts topic classification evaluation. In: *Proceedings of the 10th International Conference on Agents and Artificial Intelligence - ICAART*, vol. 2, pp. 515–522. INSTICC, SciTePress (2018)