OPTICS-based clustering of emails represented by quantitative profiles

Vladimír Špitalský and Marian Grendár

Slovanet a.s., Záhradnícka 151, 821 08 Bratislava, Slovakia
{vladimir.spitalsky,marian.grendar}@slovanet.net

Abstract. OPTICS (Ordering Points To Identify the Clustering Structure) is an algorithm for finding density-based clusters in data. We introduce an adaptive dynamical clustering algorithm based on OPTICS. The algorithm is applied to clustering emails which are represented by quantitative profiles. Performance of the algorithm is assessed on public email corpuses TREC and CEAS.

Keywords: adaptive dynamical clustering, OPTICS, DBSCAN, email clustering, spam filtering, quantitative profiles

1 Introduction

Email classification and spam filtering is a well developed area of research, which lays on an intersection of data mining, machine learning and artificial intelligence; cf. e.g. [1]. Though classification and categorization of emails have attracted considerable interest, in the literature much less attention is paid to email clustering, i.e. unsupervised learning. The main use of clustering in spam filtering is for detection of email campaigns; cf. e.g. [2, 3] or [4] for a use in semi-supervised filtering. Usually emails enter a clustering algorithm after a pre-processing, and are represented in terms of word frequencies (so-called bag-of-words representation) or by heuristic rules. Emails are commonly clustered by K-means algorithm. The bag-of-words and heuristic rules representations of emails suffer from several deficiencies such as language dependence, high computational costs, vulnerability, sensitivity to concept drift, high number of heuristic rules, necessity to update the rules, etc. Choice of the clustering algorithm is usually guided by its simplicity of use. The more advanced clustering algorithms – such as the density-based methods – are usually not considered in the email clustering context, as they require more elaborate tuning.

In [5, 6] a quantitative profile (QP) representation of emails has been proposed, and demonstrated its excellent performance in email classification and spam filtering. QP represents an email by a fixed-dimension vector of numbers that characterize the email; e.g. lengths of lines. In the present communication, the QP representation is explored in the email clustering context. It turns out that clustering by the density-based methods, such as DBSCAN and OPTICS, is well-suited for the QP representation. This motivates development of an OPTICS-based algorithm AD-OPTICS, an adaptive dynamical clustering algorithm. The main features of AD-OPTICS are: 1) multi-view clustering, 2) tuning by quantiles, 3) multi-step dynamical clustering. AD-OPTICS, when applied to categorization of QP-represented emails, gives a language independent, simple-to-use clustering algorithm, able to uncover nontrivial clusters with high homogeneity. AD-OPTICS with QPs is intended for updating a training corpus of emails, as well as a support in weighted email classification.

The paper is organized as follows. First we recall the quantitative profile representation of emails. Then the adaptive dynamical OPTICS-based clustering algorithm (AD-OPTICS) is presented in detail. Results of a performance study of AD-OPTICS in email clustering are gathered in Section 4. The concluding section summarizes the main advantages of the algorithm.

2 Quantitative profile representation of email

The Quantitative Profile (QP) approach to spam filtering and email categorization [5, 6] is motivated by the email shape analysis proposed by Sroufe, Phithakkitnukoon, Dantu, and Cangussu in [7]. Quantitative profile of an email is a p-dimensional vector of numbers that characterize the email. For instance, vector of the lengths of the first p lines of an email forms the line profile of the email. The vector of occurrences of letters from an alphabet constitutes the character profile. The line profile is an example of the binary profile, which assumes a set of special characters and the profile is obtained by counting occurrences of letters between two characters from the special set. Besides new instances of the binary profile, several other classes of quantitative profiles were considered in [6], such as the grouped character profile, d-gram grouped character profile, as well size profile. A formal definition of some of the profiles is recalled below (cf. [6] for further details).

First, note that in the QP approach, an email is represented as a realization of a vector random variable generated by a hierarchical data generating process. The length n of an email is an integer-valued random variable, with the probability distribution F_n. Given the length, the email is represented by a random vector $X_1^n = (X_1, \dots, X_n)$ from the probability distribution $F_{X_1^n \mid n}$ with the support in $\mathcal{A}^n$, where $\mathcal{A} = \{a_1, \dots, a_m\}$ is a finite set (alphabet) of size m.

The grouped character profile (CPG) is a character profile with the alphabet $\mathcal{A}'$, where $\mathcal{A}'$ is a partition of $\mathcal{A}$. Here we consider two partitions $\mathcal{A}'$ based on Unicode character categories. The CPG5 partition consists of the following five groups of characters: uppercase Lu, lowercase Ll, decimal digits Nd, controls Cc with spaces Zs, and the rest. In CPG13, the rest is further divided to punctuation Pd, Ps, Pe, Po, symbols Sm, Sc, Sk and others.

The d-gram grouped character profile (dCPG) is based on the alphabet $\mathcal{A}'^d$, where d is specified in advance. The dCPG profile consists of the counts of occurrences of the d-grams in an email.

Finally, the size profile comprises the information on the size of an email (in bytes), the size of selected header fields and also their CPGs, and the number and the size of email parts of selected content types, with their CPGs.

In [6], performance of the Random Forest-based classifiers with several classes of QP was assessed on the public TREC and CEAS email corpuses as well as on private corpuses. The performance was compared with that of the optimized SpamAssassin and Bogofilter. In the email categorization task the QP-based Random Forest classifiers attained, at much lower computational costs, comparable and even better performance than the other considered filters.

Motivated by the favorable performance of QP-based filters in the classification task, we consider their use for email clustering. The clustering is performed by a novel adaptive, dynamical, OPTICS-based clustering algorithm, which is described below.

3 Adaptive dynamical OPTICS-based clustering algorithm AD-OPTICS

Density-based clustering is an approach to clustering which utilizes density of points in forming clusters. DBSCAN and OPTICS are two of the most popular density-based clustering methods. In Section 3.1 we briefly describe these algorithms; for a survey cf. [8].

3.1 DBSCAN and OPTICS

The DBSCAN (Density-Based Spatial Clustering of Applications with Noise) clustering algorithm was proposed in [9]. It divides objects into clusters according to the given distance function d and two parameters: the radius ε and the number of points $minPts$. An object x is said to be *core* if its (closed) ε-neighborhood $U_d(x, \varepsilon)$ contains at least $minPts$ objects. The clusters are defined in such a way that they are unions of ε-neighborhoods of core points and two core points belong to the same cluster if and only if one of them is density reachable from the other one; recall that an object y is *density reachable* from x provided there are core points $x_0 = x, x_1, \ldots, x_k$ such that $x_{i+1} \in U_d(x_i, \varepsilon)$ for every i and $y \in U_d(x_k, \varepsilon)$. Objects x such that $U_d(x, \varepsilon)$ does not contain any core point are called *noise objects*; these objects do not belong to any cluster. For the details see [9].

One of the main problems with DBSCAN is the choice of the radius ε; small ε means that many objects are noise and large ε means that essentially different clusters can be glued together. To overcome these difficulties, in [10] the authors proposed OPTICS (Ordering Points To Identify the Clustering Structure) algorithm. Analogously as DBSCAN, also OPTICS depends on the distance d and two parameters $\varepsilon_{\max}$ and $minPts$. But unlike DBSCAN, OPTICS is not a clustering algorithm. Its purpose is to linearly order all objects in such a way that closest objects (according to the distance d) become neighbors in the ordering.

This is achieved by defining the so-called *core-distance* $cd(x)$ and *reachability-distance* $rd(x)$ for every object x. OPTICS guarantees that, for every $\varepsilon \leq \varepsilon_{\max}$, if $rd(x) \leq \varepsilon$ then x belongs to the same ε-DBSCAN cluster as its predecessor. Thus, for any given $\varepsilon \leq \varepsilon_{\max}$, the ε-DBSCAN clusters correspond to the maximal intervals in the OPTICS ordering such that $rd(x) \leq \varepsilon$ for every but the first object of the interval. For other algorithms extracting (hierarchical) clusterings from OPTICS, see [10–12].

Concerning the choice of $\varepsilon_{\max}$, if it is too small, OPTICS cannot extract information about clustering structure. On the other hand, with growing $\varepsilon_{\max}$ the runtime complexity of OPTICS grows dramatically, cf. e.g. [13]. Growing effort was devoted to the choice of the density threshold for DBSCAN and OPTICS, cf. e.g. [13–15]

3.2 Description of AD-OPTICS

In the proposed algorithm we address the following problems.

- First, we want the algorithm to run in a reasonable time even for large databases. Since most of the runtime is spent by OPTICS, in AD-OPTICS the value of the crucial OPTICS parameter $\varepsilon_{\max}$ is chosen in an adaptive way.
- The second problem associated with OPTICS is the issue of cluster extraction. Several methods are proposed in the literature; cf. e.g. [10–12]. We use flat DBSCAN clusters, where the cuts are determined in a novel way: by quantiles of reachability distance.
- Due to the time constraint employed to set the value of $\varepsilon_{\max}$, a part of the clustering structure can be lost. To solve this problem, the left-aside objects are sent into a new round of OPTICS.
- The final problem relates to clustering of complex objects, i.e. objects with multiple views, for which several metrics can be applied. AD-OPTICS uses a simple strategy: in each round, a metric is selected; for other multi-view density-based clustering methods, cf. e.g. [16].

Input parameters of AD-OPTICS are: the minimal cluster size *minClSize*, the minimal number *minSizeToCluster* of left-aside objects which can be further clustered and maximal time *opticsTime* for running one instance of OPTICS. The main loop of the algorithm can be summarized as follows.

1. Objects, which were left aside in the previous round, comprise the set D.
2. Select a metric d.
3. Determine $\varepsilon_{\max}$ by estimated runtime of OPTICS and *minPts* relative to given minimal cluster size.
4. Run OPTICS on D with d, $\varepsilon_{\max}$ and *minPts*.
5. Set ε-levels $(\varepsilon_i)_i$ by the quantiles of reachability.
6. Extract ε_i-DBSCAN clusters in the bottom-up way.
7. (a) If there are no clusters, go to Step 2.

(b) If the number of unclustered objects is less than *minSizeToProcess*, terminate.

(c) Otherwise, go to Step 1.

The OPTICS parameter $\varepsilon_{\max}$ is selected such that the expected OPTICS runtime is below a given time threshold *opticsTime*. An estimate of the expected runtime is based on an assumption that the expected time needed for running OPTICS is given as a product of the number of objects and the average time needed for range query with radius $\varepsilon_{\max}$.

To extract clusters from obtained OPTICS ordering, we use several DBSCAN "cuts" with ε's corresponding to quantiles of reachability distance. Notice that if ε is α-quantile of reachability distance then the ratio of clustered objects is approximately α. The obtained DBSCAN clusterings are processed from bottom to top; a cluster is accepted if it has sufficient size.

Of course, there are many other conceivable possibilities for cluster extraction. For instance, one can use a cluster validity index to accept a cluster. Further, the OPTICS reachability plot can be used for cluster extraction in various other ways, cf. e.g. [10–12]. One of the advantages of the proposed simple approach is that instead of running OPTICS on the entire dataset, the present approach allows to run OPTICS on a sample of data. Afterwards, with the OPTICS-selected ε-thresholds much simpler DBSCAN can be run on the entire dataset, saving both runtime and memory.

4 Performance study

We have applied AD-OPTICS to email clustering. To this end we have used two public corpuses TREC07 and CEAS08 of ham and spam emails. TREC07 comprises 25 220 hams and 50 199 spams, CEAS08 consists of 27 126 hams and 110 579 spams.

4.1 Implementation and configuration

The algorithm was implemented in Java and the results were processed in R [17]. The implementation of OPTICS algorithm was based on that from WEKA 3.6.5 and it was backed by Vantage Point Tree [18]. The Java library Mime4J was used for parsing emails and obtaining size profiles. Let us note that for the other four profiles (two binary and two d-gram ones) we have used no email parsing at all, with one small exception: in the two body-profiles we skipped email headers.

Emails were represented by six quantitative profiles: two size profiles (SP-CPG13 and SP-CPG5), two binary profiles (brackets profile BRP applied to the whole email and BRP applied to email body) and two d-gram grouped character profiles (3CPG5-whole and 2CPG13-body). For more detailed specification of the profiles cf. Section 2 and [6]. To every profile L_1-distance was applied. The minimal cluster size *minClSize* was set to 100 and the minimal number of objects for further clustering *minSizeToCluster* was set to 5000. The time allowed for one OPTICS run was set to 20 minutes.

4.2 Results

In this section we summarize the results obtained by AD-OPTICS on the two considered corpuses. Since the corpuses are labeled it is possible to quantify how well a clustering splits hams and spams. For every cluster we determine its *majority label*, that is, the most frequent label of emails in the cluster, and we define the *purity of a cluster* to be the percentage of emails with the majority label. The weighted (by size) average of cluster's purities is the so-called *purity of a clustering*. Formally, if $\mathcal{C}$ is a clustering and $\mathcal{L}$ is a labeling of emails, then the purity of $\mathcal{C}$ is

$$\text{purity}(\mathcal{C}) = \frac{1}{n} \sum_{C \in \mathcal{C}} \max_{L \in \mathcal{L}} |C \cap L| \ ,$$

where n is the size of the corpus.

The results are summarized by Table 1 and Figure 1. The number of produced clusters is acceptable. We can see that vast majority of emails was split into pure clusters. Among them roughly 25% were found in the first round of the algorithm. First few rounds produced almost exclusively 100% pure clusters. Naturally, the number of pure clusters, as well as purity of clusters, decreased with round.

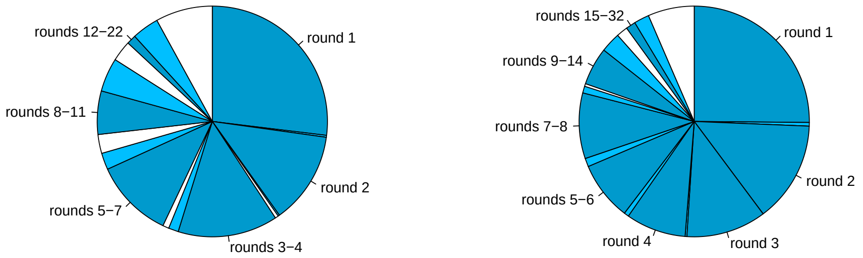
corpus	clusters	100%	90–100%	0–90%	purity
TREC07	431	331 (72.4%)	58 (12.7%)	42 (14.9%)	94.3%
CEAS08	724	638 (83.0%)	55 (8.5%)	31 (8.5%)	97.5%

Table 1: Number of clusters according to purity 100%, 90–100%, 0-90%, with size proportions in parenthesis.

Purity of clusterings obtained from AD-OPTICS is in fact the success rate of classification conditioned upon known majority labels of clusters. Hence it can be compared with success rate of supervised classification. For instance, for the TREC corpus, SpamAssassin with optimized weights has success rate 94.6%, and BogoFilter has 98.4%, cf. the supplement of [5]. From this point of view, the purity of obtained clusterings is excellent.

5 Conclusions

We have proposed an adaptive dynamical clustering OPTICS-based algorithm AD-OPTICS. The algorithm chooses the key OPTICS parameter $\varepsilon_{\max}$ in an adaptive manner and does not require laborious tuning. Clusters are extracted from OPTICS in a novel way: by quantiles of reachablity distance. Objects are clustered in a dynamic way: those which were left out of clusters, enter the new round of clustering. AD-OPTICS can naturally handle complex objects with multi-view representations. AD-OPTICS is simple to configure. Indeed, the only



(a) TREC07

(b) CEAS08

Fig. 1: Purity 100% (blue), 90–100% (light blue) and 0–90% (white) of clusters according to rounds. Later rounds are aggregated.

input parameters to the algorithm are the maximum time which a single round of the algorithm could take, the minimal cluster size, and the minimal number of the left-aside (unclustered) emails. Finally, the presented approach makes possible running OPTICS on a sample of data, and then, using the OPTICS-suggested ε-thresholds, apply DBSCAN clustering on the whole dataset.

The algorithm was applied to email clustering, with emails represented by several Quantitative Profiles (QP). The AD-OPTICS configuration parameters are independent of the particular QP used for email representation. On two public corpuses AD-OPTICS produced excellent results. Vast majority of emails were clustered into homogeneous clusters. Moreover, along “trivial” clusters (those, for which one can easily construct a regular expression for filtering them out), also highly non-trivial and still homogeneous clusters were obtained.

The AD-OPTICS application to clustering QP-represented emails is easy to extend in several directions: emails can be represented by any QP and distance of emails can be measured by different metrics. When predicted labels of emails are available, they can be seamlessly integrated into the clustering process. Thanks to the QP representation, language of email does not matter.

Acknowledgments. Stimulating feedback from Peter Farkaš, Ján Gallo, Peter Kočík, Erik Lehotský, Jana Majerová, Dušan Slivoň, Jana Škutová and Stanislav Záriš is gratefully acknowledged. This paper was prepared as a part of the “SPAMIA” project, MŠ SR 3709/2010-11, supported by the Ministry of Education, Science, Research and Sport of the Slovak Republic, under the heading of the state budget support for research and development.

References

1. Almeida, T.A., Yamakami, A.: Advances in spam filtering techniques. In: Studies in Computational Intelligence, 394, pp. 199–214. Springer, Heidelberg (2012)

2. Haider, P., Scheffer, T.: Bayesian clustering for email campaign detection. In: ICML'09, pp. 385–392. ACM, New York (2009)
3. Qian, F., Pathak, A., Charlie Hu, Y., Morley Mao, Z., Xie, Y.: A case for unsupervised-learning-based spam filtering. In: SIGMETRICS'10, pp. 367–368. ACM, New York (2010)
4. Whissell, J.S., Clarke, Ch.L.A.: Clustering for semi-supervised spam filtering. In: CEAS'11, pp. 125–134. ACM, New York (2011)
5. Grendár, M., Škutová, J., Špitalský, V.: Spam filtering by quantitative profiles. Intnl. J. Comp. Sci. Issues, 9, 265–271 (2012)
6. Grendár, M., Škutová, J., Špitalský, V.: Email categorization and spam fitering by random forest with new classes of quantitative profiles. In: Compstat 2012, pp. 283–294. ISI/IASC (2012)
7. Sroufe, P., Phithakkitnukoon, S., Dantu, R., Cangussu, J.: Email shape analysis. In: Distributed Computing and Networking, LNCS 5935, pp. 18–29. Springer, Heidelberg (2010)
8. Kriegel, H.P., Kröger, P., Sander, J., Zimek, A.: Density-based clustering. WIREs DMKD 1(3), 231-240 (2011)
9. Ester, M., Kriegel, H.P., Sander, J., Xu, X.: A Density- Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. In: Proc. 2nd Int. Conf. on KDDM, pp. 226–231. AAAI Press (1996)
10. Ankerst, M., Breunig, M., Kriegel, H.P., Sander, J.: OPTICS: Ordering Points To Identify the Clustering Structure. In: Proc. ACM SIGMOD 99, pp. 49–60. Springer, Heidelberg (1999)
11. Sander, J., Qin, X., Lu, Z., Niu, N., Kovarsky, A.: Automatic Extraction of Clusters from Hierarchical Clustering Representations. In: Proc. 7th Pacific-Asia Conference on KDDM, pp. 75-87 (2003)
12. Brecheisen, S., Kriegel, H.P., Kröger, P., Pfeifle, M.: Visually Mining Through Cluster Hierarchies. In: Proceedings of the 4th SIAM International Conference on Data Mining, pp. 400-411. SIAM (2004)
13. Achtert, E., Böhm, C., Kröger, P.: DeLi-Clu: Boosting Robustness, Completeness, Usability, and Efficiency of Hierarchical Clustering by a Closest Pair Ranking. In: Advances in KDDM, LNCS 3918, pp. 119-128. Springer, Heidelberg (2006)
14. Gorawski, M., Malczok, R.: AEC Algorithm: A Heuristic Approach to Calculating Density-Based Clustering Eps Parameter. In: ADVIS 2006, LNCS 4243, pp. 90–99. Springer, Heidelberg (2006)
15. Cassisi, C., Ferro, A., Giugno, R., Pigola, G., Pulvirenti A.: Enhancing density-based clustering: Parameter reduction and outlier detection. Info. Sys. 38, 317–330 (2013)
16. Achtert, E., Kriegel, H.P., Pryakhin, A., Schubert, M.: Hierarchical Density-Based Clustering for Multi-Represented Objects. In: MCD'05, ICDM (2005)
17. R Development Core Team: R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0. `http://www.R-project.org` (2010)
18. Yianilos, P.N.: Data structures and algorithms for nearest neighbor search in general metric spaces. In: Proceedings of the fourth annual ACM-SIAM Symposium on Discrete algorithms, pp. 311-321. SIAM (1993)