Topic detection using the DBSCAN-Martingale and the Time Operator

Ilias Gialampoukidis[1,2], Stefanos Vrochidis[2], Ioannis Kompatsiaris[2], and Ioannis Antoniou[1]

[1] Department of Mathematics, Aristotle University of Thessaloniki
54124 Thessaloniki, Greece
(E-mail: iliasfg@math.auth.gr; iantonio@math.auth.gr)
[2] Information Technologies Institute, Centre for Research and Technology-Hellas
6^{th} km Charilaou-Thermi road, 57001 Thessaloniki, Greece
(E-mail: heliasgj@iti.gr; stefanos@iti.gr; ikom@iti.gr)

Abstract. Topic detection is usually considered as a decision process implemented in some relevant context, for example clustering. In this case, clusters correspond to topics that should be identified. Density-based clustering, for example, uses only a density level ϵ and a lower bound for the number of points in a cluster. As the density level is hard to be estimated, a stochastic process, called the DBSCAN-Martingale, is constructed for the combination of several outputs of DBSCAN for various randomly selected values of ϵ in a predefined closed interval $[0, \epsilon_{max}]$ from the uniform distribution. We have observed that most of the clusters are extracted in the interval $[0, \epsilon_{max}/2]$, and moreover in the interval $[\epsilon_{max}/2, \epsilon_{max}]$ the DBSCAN-Martingale stochastic process is less innovative, i.e. extracts only a few or no clusters. Therefore, non-symmetric skewed distributions are needed to generate density levels for the extraction of all clusters in a fast way. In this work we show that skewed distributions may be used instead of the uniform, so as to extract all clusters as quickly as possible. Experiments on real datasets show that the average innovation time of the DBSCAN-Martingale stochastic process is reduced when skewed distributions are employed, so less time is needed to extract all clusters.
Keywords: DBSCAN-Martingale, Time Operator, Skewed distributions, Internal Age, Density-based Clustering, Innovation process.

1 Introduction

Journalists and media monitoring companies need to quickly detect interesting articles of the same topic and to manage large collections of news articles. Given a collection of news articles the estimation of the correct number of topics is a challenging task, due to the fact that there are news articles that do not belong to any of the topics. We have presented an estimation on the number of clusters (topics) using a Martingale process, namely the DBSCAN-Martingale [1]. The DBSCAN [2] algorithm is repeatedly applied using a random density parameter ϵ, while the lower bound for the number of clusters $minPts$ is kept constant.

17^{th} *ASMDA Conference Proceedings, 6 – 9 June 2017, London, UK*

The generated stochastic process progressively estimates the number of clusters in any dataset but has been introduced in the context of text clustering to estimate the number of topics. The final number of clusters is provided by a majority vote among several realizations of the DBSCAN-Martingale process. Similarly, the DBSCAN-Martingale has also been applied in the context of image retrieval and image clustering [3] in the estimation of the number of visual words in a set of visual descriptors, showing the wide applicability of this novel clustering approach. In all cases, the processing time is a critical aspect and needs to be minimized as much as possible.

Towards this direction, we examine whether skewed distributions are able to extract all clusters faster than the uniform distribution in the selection of the density level ϵ. The time needed to extract all clusters is based on the number of iterations, using several random choices of the density level ϵ in a pre-defined interval $[0, \epsilon_{max}]$. However, not all iterations of DBSCAN are innovative, i.e. they do not extract the same number of clusters or some iterations do not extract any clusters. The innovation probabilities at any stage of a stochastic process have been introduced in [4] and have been demonstrated in the non-stationary random walks modeling stock market prices [5] and in the fluctuations of the US economy [6], through the construction of the associated Time Operator. The Time Operator has been introduced in the context of stochastic processes [7–9], quantifying the distribution of innovations in the considered (clock) time domain. We shall examine whether the innovations of the DBSCAN-Martingale are distributed in a symmetric way or not, in order to minimize the required time stages T needed to extract all clusters (topics).

2 Density-based clustering

DBSCAN [2] provides as output a clustering vector C with values the cluster IDs $C[k]$ of each point $k = 1, 2, \ldots, n$, assigning each item k to a cluster. In case the k-th item is marked as noise, then: $C[k] = 0$. The parameters of DBSCAN are, first, a density level ϵ and, second, a lower bound for the number of clusters in a dataset $minPts$. The parameter $minPts$ is usually predefined based on the size of the expected clusters, but the density level ϵ is hard to be estimated and, if so, then the algorithm is not able to output all clusters using one unique density level, as shown, for example, in Fig. 1, where there are 10 clusters, but not of the same density level.

The OPTICS diagram [10] has been used to visualize the cluster structure, where each dent represents a cluster. Moreover, OPTICS is used to observe the density level at which the desired clusters are extracted. The OPTICS plot for the dataset of Fig. 1 is presented in Fig. 2.

The parameter $minPts$ is a pre-defined fixed value, approximately equal to 10, as initially proposed in [10]. For each density level ϵ, the output of DBSCAN is one clustering vector and is denoted by $C_{DBSCAN(\epsilon)}$. Small values of ϵ result to $C_{DBSCAN(\epsilon)} = \mathbf{0}$, where $\mathbf{0}$ is a vector of zeros, because all points are marked as noise. However, large values of ϵ, result to $C_{DBSCAN(\epsilon)} = \mathbf{1}$, where $\mathbf{1}$ is a vector of ones, since all points are reachable from any other point, hence, all points are assigned to the same cluster. A smart selection of the density

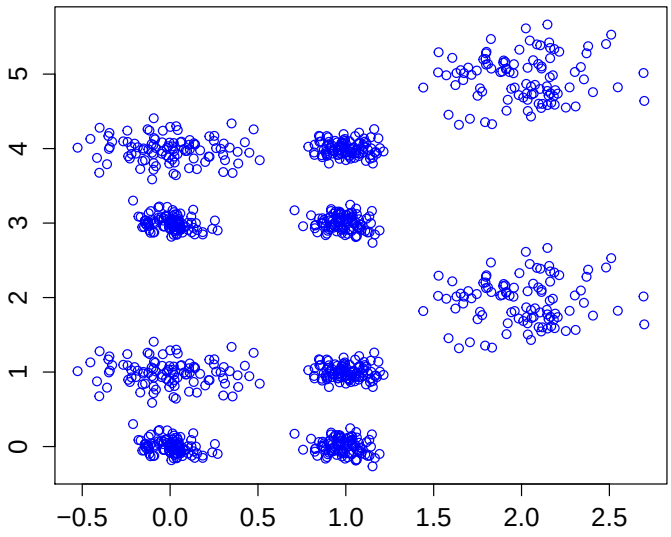

Fig. 1. A dataset with 10 clusters

level cannot ensure correct estimation in the number of clusters, with a strong impact to the performance of a news clustering approach [1], using for example Latent Dirichlet Allocation [11] to assign news articles to topics.

The estimated number of clusters in the illustrative dataset of Fig. 1 is presented in Fig. 3, where we observe that 10 clusters are correctly estimated by the majority of 1000 DBSCAN-Martingale realizations.

3 The DBSCAN-Martingale and Time Operator

Initially, a random sample of uniformly distributed random numbers $\epsilon_t, t = 1, 2, \ldots, T$ in $[0, \epsilon_{max}]$ is generated by the DBSCAN-Martingale. The sample of $\epsilon_t, t = 1, 2, \ldots, T$ is sorted in increasing order and for each density level ϵ_t a clustering vector is provided by DBSCAN, denoted by $C_{DBSCAN(\epsilon_t)}$.

In the beginning of the DBSCAN-Martingale process, there are no clusters detected, i.e. $C^{(0)} = \mathbf{0}$. We denote by $\mathcal{F}_t$ the σ-algebra generated by $\{C_{DBSCAN(\epsilon_1)}, C_{DBSCAN(\epsilon_2)}, \ldots, C_{DBSCAN(\epsilon_t)}\}$ and let $\mathcal{F}_0$ be the trivial σ-algebra $\{\Omega, \emptyset\}$ at stage $t = 0$. At stage $t = 1$ all clusters are kept: $C^{(1)} := C_{DBSCAN(\epsilon_1)}$, extracted at lowest density level ϵ_1. At stage, $t = 2$, some of the detected clusters by $C_{DBSCAN(\epsilon_2)}$ are new and some of them have also been extracted at stage $t = 1$. DBSCAN-Martingale keeps only the newly detected clusters of the second stage, $t = 2$, by taking only groups of points of

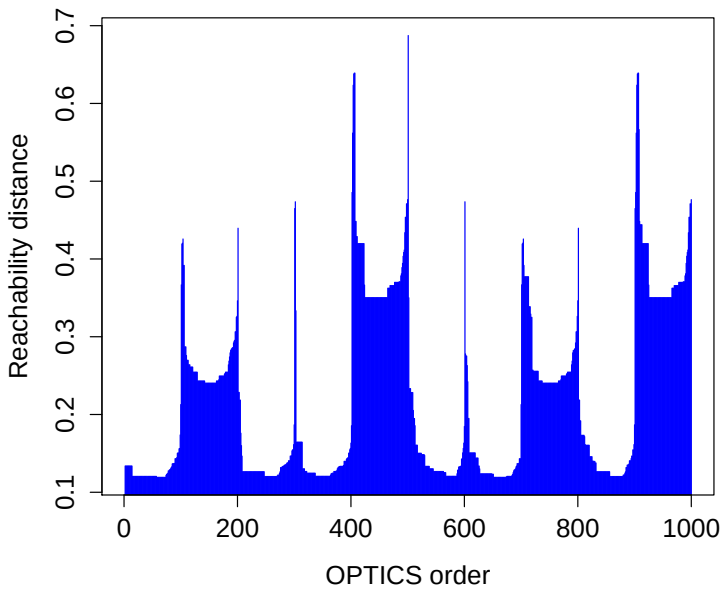

Fig. 2. OPTICS reachability distance plot for the dataset of Fig. 1

the same cluster ID with size greater than $minPts$:

$$C^{(t)}[j] := \begin{cases} 0 & \text{if point } j \text{ belongs to a previously extracted cluster} \\ C_{DBSCAN(\epsilon_t)}[j] & \text{otherwise} \end{cases} \tag{1}$$

where $C^{(1)} = C_{DBSCAN(\epsilon_1)}$. Each vector of Eq. (1) has only the newly extracted clusters and all other points are marked as noise (zero cluster ID). The cluster IDs of $C^{(t)}$ are relabeled, starting from $1 + \max_j C^{(t-1)}[j]$ to $r + \max_j C^{(t-1)}[j]$, assuming that r clusters have been extracted up to stage t.

The Hilbert space $\mathcal{H}_t$ is generated by the conditional expectations $\mathbb{E}_t = E[.|\mathcal{F}_t]$ up to stage $t, t = 1, 2, \ldots, T$, where the σ-algebras $\mathcal{F}_t, t = 1, 2, \ldots, T$ are generated by the vectors $\{C_{DBSCAN(\epsilon_1)}, C_{DBSCAN(\epsilon_2)}, \ldots, C_{DBSCAN(\epsilon_t)}\}$. Our knowledge about the final clustering vector C up to stage t is restricted to $\mathbb{E}_t C$. Moreover, the projections onto the innovation spaces $\mathcal{N}_t$ are given by:

$$\mathbb{P}_t C = E[C|\mathcal{F}_t] \ominus E[C|\mathcal{F}_{t-1}] = (\mathbb{E}_t \ominus \mathbb{E}_{t-1})C = C^{(t)} \tag{2}$$

and the final clustering vector C lives in the space of fluctuations $\mathcal{H} = \mathcal{N}_1 \oplus \mathcal{N}_2 \oplus \cdots \mathcal{N}_T$:

$$C = C^{(1)} \oplus C^{(2)} \oplus \ldots \oplus C^{(T)} \tag{3}$$

The projections $\mathbb{E}_t C = E[C|\mathcal{F}_t], t = 1, 2, \ldots, T$ are our "best prediction" about the final outcome of the clustering vector C which needs to be determined:

$$\mathbb{E}_t C = E[C|\mathcal{F}_t] = C^{(1)} \oplus C^{(2)} \oplus \ldots \oplus C^{(t)} \tag{4}$$

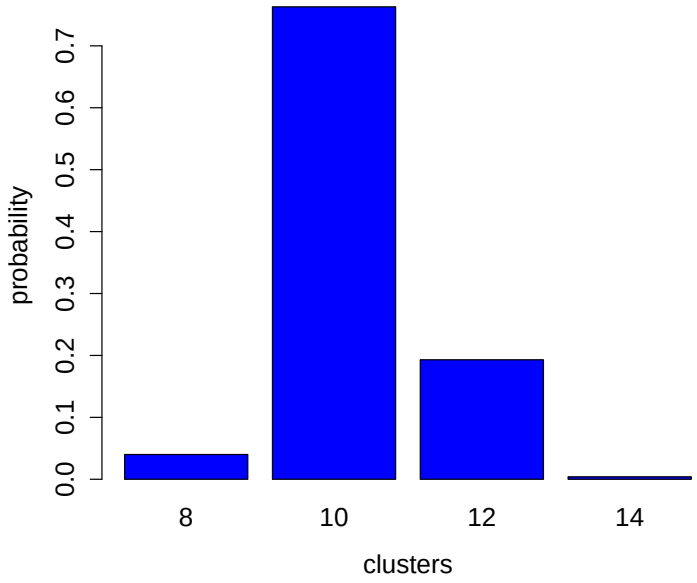

Fig. 3. 1000 realizations of the DBSCAN-Martingale in the dataset of Fig. 1

Finally, at stage $t = T$, we have gained all available knowledge about the vector C, i.e. $C = E[C|C_{DBSCAN(\epsilon_1)}, C_{DBSCAN(\epsilon_2)}, \ldots, C_{DBSCAN(\epsilon_T)}]$ and all available clusters have been extracted.

The self-adjoint operator with spectral projections the conditional expectations $\mathbb{E}_t$ on the space of fluctuations $\mathcal{H}$ is the *Time Operator* of the stochastic process $X_t, t = 1, 2, \ldots$:

$$\mathbb{T} = \sum_{t=1}^{\infty} t(\mathbb{E}_t \ominus \mathbb{E}_{t-1}) \tag{5}$$

The Time Operator, as defined in Eq. (5), acts on the clustering vector C in $\mathcal{H}$, defining also the distribution of innovations:

$$p_t = \text{Prob}\{C \in \mathcal{N}_t\} = \frac{\|\mathbb{P}_t C\|^2}{\|C - E[C]\|^2} = \frac{\|C^{(t)}\|^2}{\|C\|^2} \tag{6}$$

where $E[C] = 0$ because at the beginning of the process the clustering vector C is a vector of zeros and there are no expected clusters without any application of the DBSCAN algorithm.

The distribution of innovations has been assumed to be symmetric in [1], since the random sample of density levels $\epsilon_t, t = 1, 2, \ldots, T$ in $[0, \epsilon_{max}]$ has been generated from the uniform distribution. In contrast, we propose the generation of the random sample in an alternative way, having statistically significant skewness.

4 Generation of skewed samples of density levels

Motivated by the generation of random samples from the exponential distribution [12], through the transformation

$$X \leftarrow -\frac{\ln U}{\lambda}$$

where U is a random variable uniformly distributed in [0,1], we propose the following generation of a random sample of density levels as follows:

1. Generate a sample of size T from the uniform distribution in [0,1]:

$$\epsilon_1, \epsilon_2, \ldots, \epsilon_T$$

2. Transform the generated values using the natural logarithm:

$$\epsilon_t \leftarrow -\ln \epsilon_t$$

3. Normalize in [0,1]:

$$\epsilon_t \leftarrow \epsilon_t / \max_t \{\epsilon_t\}$$

4. Expand in $[0, \epsilon_{max}]$:

$$\epsilon_t \leftarrow \epsilon_t * \epsilon_{max}$$

This generation of the sample is parameter free (no rate parameter λ is required) and is in fact skewed, since it is a normalized sample of the exponential distribution, which in general has skewness equal to two. The sample skewness is usually estimated in three different ways [13], as also highlighted in the documentation of the library "e1071" of the statistical software R. We selected the typical definition used in many textbooks:

$$g_1 = \frac{m_3}{m_2^{3/2}} \tag{7}$$

where the sample moments of order r are:

$$m_r = \frac{1}{n} \sum_{i=1}^{n} (x_i - \mu)^r \tag{8}$$

and x_i are the non-missing elements of x, μ their mean value.

In the following, we shall examine whether the proposed steps 1–4 reduce the innovation time of the DBSCAN-Martingale process, by reducing the stages T needed to extract all clusters.

---

https://www.r-project.org/

5 Experiments

In the experiments we compare our proposed method with the uniform generation of density levels ϵ. The datasets we have used for comparison are four synthetic datasets with points in the 2-dimensional plane that contain 5,10, 15 and 20 clusters, with sizes 500, 1000, 1500 and 2000 points, respectively. We also downloaded the news articles available in the datasets WikiRef150, WikiRef186 and WikiRef220, containing, topics such as Paris attacks November 2015, Premier League, Malaysia Airlines Flight 370, Samsung Galaxy S5 and Michelle Obama (5 topics). Additional information is provided in the online dataset description. We used the online implementation of DBSCAN-Martingale.

Table 1. Time needed to extract all clusters using the uniform distribution as proposed in [1] and using our proposed skewed sample. In bold we present the minimum values for the time needed to extract all clusters.

Realizations		Uniform distribution		Skewed distribution	
Dataset	Clusters	Skewness	time needed T	Skewness	time needed T
Dataset 1	5	0.02	4	1.12	**2**
Dataset 2	10	-0.01	4	0.99	**2**
Dataset 3	15	0.00	3	1.02	**2**
Dataset 4	20	-0.02	3	0.97	**2**
WikiRef150	3	0.01	**3**	0.99	**3**
WikiRef186	4	0.00	4	0.93	**3**
WikiRef220	5	0.00	4	0.96	**4**

In Table 1 we observe that the time needed to extract all clusters is significantly reduced by our approach. This fact has a strong impact in the overall estimation of the number of clusters or topics, since the DBSCAN-Martingale process is generated several times and the final decision is taken by a majority vote scheme. Apparently, there are cases, such as the WikiRef150, where the clusters are 3 and both methods extract the clusters using the same time.

6 Concluding Remarks

We have presented a novel approach to generate the sample of density levels in the density-based clustering approach of DBSCAN-Martingale. The innovation time, as also modeled by the associated Time Operator, is reduced when skewed non-symmetric samples are employed, in all datasets examined. The proposed approach has been tested in three datasets of news articles and in four general synthetic datasets with various sizes and numbers of clusters. The skewed generation of the density levels is able to reduce the time needed to extract all clusters and therefore, provides faster estimation of the number of clusters. In the future, we shall examine whether this approach is also applicable to other clustering tasks in multimedia and social media applications.

http://mklab.iti.gr/project/web-news-article-dataset
https://github.com/MKLab-ITI/topic-detection/blob/master/DBSCAN_Martingale.r

Acknowledgements

The first author would like to thank the Research Committee of the Aristotle University of Thessaloniki for awarding him the "Aristeia" postdoctoral scholarship 2016. Moreover, this work has been partially supported by the EC-funded project KRISTINA (H2020-645012).

References

1. Ilias Gialampoukidis, Stefanos Vrochidis, and Ioannis Kompatsiaris. A hybrid framework for news clustering based on the dbscan-martingale and lda. In *Machine Learning and Data Mining in Pattern Recognition*, pages 170–184. Springer, 2016.
2. Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu, et al. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Kdd*, volume 96, pages 226–231, 1996.
3. Ilias Gialampoukidis, Stefanos Vrochidis, and Ioannis Kompatsiaris. Incremental estimation of visual vocabulary size for image retrieval. In *INNS Conference on Big Data*, pages 29–38. Springer, 2016.
4. Ilias Gialampoukidis, Karl Gustafson, and Ioannis Antoniou. Financial time operator for random walk markets. *Chaos, Solitons & Fractals*, 57:62–72, 2013.
5. Ilias Gialampoukidis and Ioannis Antoniou. Time operator and innovation. applications to financial data. In Lidia Filus, Teresa Oliveira, and Christos H Skiadas, editors, *Stochastic Modeling Data Analysis & Statistical Applications*, chapter 1, pages 19–31. ISAST, 2015.
6. Ilias Gialampoukidis, Karl Gustafson, and Ioannis Antoniou. Time operator of markov chains and mixing times. applications to financial data. *Physica A: Statistical Mechanics and its Applications*, 415:141–155, 2014.
7. Ioannis Antoniou and Karl Gustafson. Wavelets and stochastic processes. *Mathematics and Computers in Simulation*, 49(1):81–104, 1999.
8. Karl E Gustafson. *Lectures on computational fluid dynamics, mathematical physics, and linear algebra*. World Scientific, 1997.
9. Ioannis Antoniou and Theodoros Christidis. Bergson's time and the time operator. *Mind and Matter*, 8(2):185–202, 2010.
10. Mihael Ankerst, Markus M Breunig, Hans-Peter Kriegel, and Jörg Sander. Optics: ordering points to identify the clustering structure. In *ACM Sigmod Record*, volume 28, pages 49–60. ACM, 1999.
11. David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022, 2003.
12. Luc Devroye. Sample-based non-uniform random variate generation. In *Proceedings of the 18th conference on Winter simulation*, pages 260–265. ACM, 1986.
13. DN Joanes and CA Gill. Comparing measures of sample skewness and kurtosis. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 47(1):183–189, 1998.