ACTIVE LEARNING FOR SEMI -SUPERVISED DENSITY BASED CLUSTERING

Vu Viet Vu[*]
College of Technology - TNU

SUMMARY

The active learning problem for semi-supervised clustering is an active topic for the last ten years. The aim of this paper is to propose a method that is able to collect the labeled data (called seed) to improve the quality of seed based clustering algorithms and reduce the questions to experts. To do this task, we use the k-nearest neighbor graph to express input data and apply a local density function to evaluate the density of each data point. Then, the points that are in the dense regions will be chosen to get label by experts. Our experimental results according to our method when compared with other algorithms present its own benefits.

Key words: clustering, semi-supervised clustering, active learning, seeds

INTRODUCTION

In recent years, semi-supervised clustering algorithms using the side information (seed or pairwise constraints) have attracted a lot of attention from the machine learning community, as they promise to improve the quality of traditional methods [8,9].

Active learning provides an efficient way for semi-supervised clustering algorithms to retrieve the side information they rely on: the algorithm asks the expert for the value of a class label or a relationship between instances.

This paper specifically focuses on an active seed selection algorithm that queries the expert to retrieve class labels. The researcher conducted in the field which mainly focused on adapting well-known clustering methods to this new semi-supervised context. In additions, we particularly aim at guiding the exploration of the space searching to relevant solutions, or overcoming some inherent limitations of clustering algorithms. For example, seed k-means (SKM) or seed fuzzy c-means (SFCM) [2, 10] allows us to reduce the sensitivity of these methods to their initial partition. Similarly, seeds have been used to estimate distinct local density parameters in density-based algorithms like SSDBSCAN [11]. However, all these methods do not address the problem of how to select the most appropriate seeds for their needs: whereas a number of researches have been conducted in the context of semi-supervised classification [12], just few methods have been proposed in the clustering context. Moreover, the existing methods are limited by hypothesis on the underlying data distribution and on the shape and sizes of expected clusters [2, 7].

To this aim, this paper introduces a new efficient algorithm for active seeds selection, that can adapt with any seed-based clustering algorithm, and that relies directly on a k-nearest neighbors graph to identify the regions of data space in which requesting the expert for labeled instances.

This paper is organized as follows: Section 2 reviews the main active seed-selection methods. Then, Section 3 introduces our new active seed selection method based on a k-nearest neighbors graph. Section 4 describes the experiments. Finally, Section 5 presents the conclusions and perspectives of this research.

RELATED WORK

The problem of selecting the best seeds in the context of clustering algorithms has already been partially covered by papers related to the problem of initialization of centers in k-means like algorithms [2]. As recalled by [2], this problem has been deeply studied but one can

[*] *Tel: 0986 439559, Email: vuvietvu@tnut.edu.vn*

identify four major approaches to initialize the centers in k-means like approaches: the random creation of the initial partition, the classical Forgy method as reported by [3] in which initial seeds are randomly selected (and then all points are assigned to the nearest seed), the MacQueen method in which, similarly to Forgy, seeds are chosen randomly, but then each time a point is assigned to a seed, the corresponding cluster center is updated, and finally the Kaufman approach [4] in which the first seed is the center of the data set and all the other seeds are selected according to a criterion that depends on the number of data in the neighborhood of the seed candidate and the distance to the seeds that are already selected.

More precisely, in [5], the author proposes two heuristics that maximize either the sum of the distance or the minimal distance to the existing seeds. Finally, the other works like [6] propose to initialize the first seed as the center of the dataset and then to randomly select the other points that are averaged with the center coordinate with an appropriate weight to cover the entire dataset while being more resistant to outliers.

All these methods (random selection or maximization of the distance to existing seeds) allow an efficient coverage of the data space and some of the approaches like [4] also take into account a density measure (number of data points in the neighborhood) to choose from all the possible distant seeds.

The next sections briefly describe the main Min-Max method that has been proposed to allow the active selection of seeds by an expert in the context of semi-supervised clustering [7].

MIN-MAX APPROACH

The objective of the Min-Max approach (SMM) is to build a set of seeds Y from a data set X such as the seeds in Y are evenly spaced and produce a good coverage of the data space [7]. Moreover, the method aims at minimizing the annotation effort of the expert and thus tries to minimize the number of seeds that represent the same cluster. Initially, as there is a prior information about the data set, the first seed of the set Y is randomly chosen among data points in X. Then, the next seeds have to maximize their minimal distance to the set of seeds which has been already selected as shown in the following Equation:

$$y_{new} = \mathrm{argmax}_{x\in X}(\mathrm{min}_{y\in Y}\mathrm{d}(x,y)) \qquad (1)$$

where y_{new} denotes the new point added to the seeds set Y and where d(.) denotes a metric defined in the data space of points X (for example d(.) could be an Euclidian distance or a Mahanalobis distance if we compare R^m vectors, a Levenshtein distance if we compare sequences...

The active seed selection algorithm based on the Min-Max approach is an iterative process where a new seed candidate y_{new} at each step (as determined by Equation 1) is proposed to the expert to be labeled. In any active learning system, the expert is supposed to be able to answer to all the queries of the system. The iterative process stops when the experts make a decision or when all of the points in X have been explored.

One shortcoming of this method is the results which strongly depend on the first selected seed, the shape and the size of the clusters and the number of seeds available for each cluster.

PROPOSED METHOD

To develop the new active learning method, we use a k-nearest neighbor graph which is introduced in [13]. The k-nearest neighbor graph is a weighted undirect graph, in which each vertex represents a data point, and possesses at most k edges to its k-nearest neighbors. An edge is created between a pair of vertices, u and v, if and only if the points associated to vertices u and v have each other in their k-nearest neighbors set. The weight ω(u, v) of the edge between the vertices u and v is defined as the number of common nearest neighbors the two points associated to u and v share, as shown in equation that is expressed as followings:

$\omega(x_i, x_j) = |NN(x_i) \cap NN(x_j)|$

where NN(.) denotes the set of k-nearest neighbors of the associated point. The important property of this similarity measure is its own built-in automatic scaling, which makes it adapted to treat datasets with distinct cluster densities. Figure 1, 2 and 3 show examples about the k-nearest neighbor graph.

Then, it is possible to extract a local density indicator from a k-NNG with the Local Density Score [14], which is defined as the average, for each point, of the proximity ω with all its neighbors as recalled in Equation as follows.

$$LDS(x) = \frac{\sum_{q \in NN(x)} \omega(x,q)}{k}$$

The LDS value of a point x is set in [0, k-1] where k is the number of the nearest neighbors. It is defined such as a high value of LDS(x) which indicates a high proximity between the point x and its neighbors, i.e. x belongs to dense region of the data space. Similarly, a small value of LDS(x) indicates that x belongs to a transition region between clusters or x is an outlier with far nearest neighbors.

The new proposed method, called SkNN-LDS, focuses on candidate seeds near the centers of potential clusters, and the candidate seeds *Candidate_set* is as follows.

Candidate_set = $\{p \in X: LDS(x) \geq \varepsilon\}$

Using the *Candidate_set*, we will construct all the connected components and after we sort these connected components in the descending of vertices number. At each step, the connected component with maximum of vetice will be chosen for getting its label. The query process will stop by user or when the Candidate_set is empty.

The algorithm SkNN-LDS is detailed in Algorithm 1.

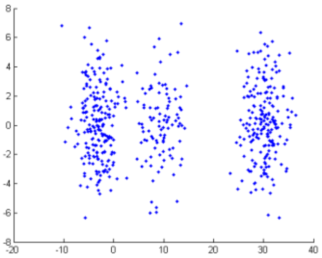

Figure 1. *Data set*

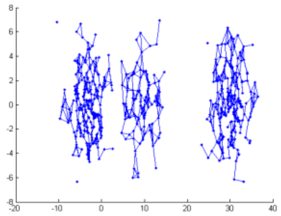

Figure 2. *7-Nearest Neighbor Graph*

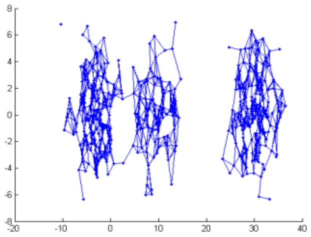

Figure 3. *10-nearest neigbor graph*

```
Algorithm 1: SkNN-LDS
Input: Data set X, k, and $\varepsilon$
Output: Set of seed Y
Begin
Y = $\Phi$;
Construct the kNN graph
Calculate the LDS value for all u $\in$ X
C = {u $\in$ X: LDS(u) $\geq \varepsilon$ }
Build the set of connected components from C:
CC = {$C_1$, $C_2$, …, $C_m$}
Repeat
Randomly select u $\in C_v$ such that
$|C_v|=\max_{c\in CC}|c|$
Query the expert to get the class label of u
 Y = Y $\cup$ {$C_v$}
  CC = CC – {$C_v$}
Until (CC = $\varnothing$) or (User_stop = true)
Return Y
End
```

EXPERIMENTAL RESULTS

We use 6 real datasets from the Machine Learning Repository [15] to evaluate my algorithm. The detail of datasets is shown in Table 1.

Table 1. *Data set for testing*

ID	Name	N	M	K
1	Protein	115	20	6
2	Iris	150	4	3
3	Glass	214	9	6
4	Thyroid	215	5	3
5	LetterIJL	227	16	3
6	Zoo	112	3	6

We use the Rand Index (RI) measure [13], as it is widely used in evaluation of clustering results. We compare our method with SMM and S-Random algorithms. The results are showed in Figure 4.

It can be seen from Figure 4 that SkNN-LDS outperforms SMM and S-Random for each of the benchmark data sets. It can be explained that the k-NN graph is very adapt to express data. Moreover, by using LDS function, we can evaluate the local density score for each data point before choosing to get its label and hence the result of clustering is increased.

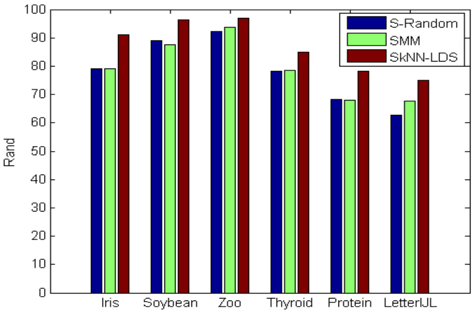

Figure 4. *Experiment results*

CONCLUSION

This paper presents a new active learning method for semi-supervised clustering. By using the k-NN graph, we developed an efficient algorithm to collect the seed that can boost the quality of semi-supervised clustering. In future works, we will apply this research for some real applications such as image/speech processing.

REFERENCE

1. J.M. Penna, J.A. Lozano, and P. Larranaga. An empirical comparison of four initialization methods for the k-means algorithm. Pattern Recognition Letters, 20:1027–1040, 1999.

2. Amine Ben said, Lawrence O. Hall, James C. Bezdek, Laurence P. Clarke: Partially supervised clustering for image segmentation. Pattern Recognition 29(5): 859-871, 1996.

3. M.R. Anderberg. Cluster Analysis for Applications. Academic Press, New York, 1973

4. L. Kaufman and P.J. Rousseeuw. Finding groups in data: An introduction to cluster analysis. In John Wiley and Sons, 1990.

5. M. Snarey, N.K. Terrett, P. Willet, and D.J. Wilton. Comparison of algorithms for dissimilarity-based compound selection. J.Mol. Graphics and Modelling, 15: 372-385, 1997.

6. J. Heer and E. Chi. Identification of web user trffic composition using multi-modal clustering and information scent. In proc. Of the workshop on web mining, SIAM conference on Data mining, 2001.

7. Viet-Vu Vu, Nicolas Labroche, and Bernadette Bouchon-Meunier. Active Learning for Semi-Supervised K-Means Clustering. In Proceedings of the 22nd IEEE International Conference on Tools with Artificial Intelligence (ICTAI-2010), Arras, France, October, 2010.

8. Anil K. Jain: Data clustering: 50 years beyond K-means. Pattern Recognition Letters (PRL) 31(8):651-666 (2010).

9. S. Basu, I. Davidson, and K. L. Wagstaff, Constrained Clustering: Advances in Algorithms, Theory, and Applications, Chapman and Hall/CRC Data Mining and Knowledge Discovery Series, 1st edn., 2008.

10. Sugato Basu, Arindam Banerjee, Raymond J. Mooney: Semi-supervised Clustering by Seeding. ICML 2002: 27-34, 2002.

11. Levi Lelis, Jörg Sander: Semi-supervised Density-Based Clustering. ICDM : 842-847, 2009.

12. B. Settles, Active Learning Literature Survey, Technical Report 1648, University of Wisconsin-Madison, 2010.

13. R.A. Jarvis and E.~A. Patrick, Clustering using a similarity measure based on shared near neighbors, IEEE Transactions on Computer, 22(11), pp: 1025-2034, 1973.

14. Duy-Dinh Le, Shin'ichi Satoh: Unsupervised Face Annotation by Mining the Web. ICDM 2008: 383-392.

15. M. Lichman, UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science, 2013.

16. W. M. Rand. Objective criteria for the evaluation of clustering methods. Journal of the American Statistical Association, 66 (336): 846–850, 1971.

TÓM TẮT
NGHIÊN CỨU PHƯƠNG PHÁP HỌC TÍCH CỰC CHO BÀI TOÁN PHÂN CỤM NỬA GIÁM SÁT DỰA TRÊN MẬT ĐỘ

Vũ Việt Vũ[*]
Trường Đại học Kỹ thuật Công nghiệp – ĐH Thái Nguyên

Vấn đề học tích cực sử dụng cho bài toán phân cụm nửa giám sát là một trong những chủ đề thu hút được nhiều sự quan tâm trong khoảng mười năm trở lại đây. Trong bài báo này, chúng tôi đề xuất một thuật toán nhằm thu thập các seed (labeled data) với mục tiêu làm tăng chất lượng của các thuật toán phân cụm nửa giám sát và đồng thời giảm số câu hỏi đối với các chuyên gia trong lĩnh vực trong quá trình thu thập các seed. Thuật toán được xây dựng dựa trên đồ thị k-láng giềng gần nhất và một hàm đánh giá mật độ của các điểm trên đồ thị. Kết quả thực nghiệm cho thuật toán đề xuất đạt kết quả tốt hơn so với các thuật toán cùng loại.

Key words: *clustering, semi-supervised clustering, active learning, seeds*

Ngày nhận bài:20/6/2015; Ngày phản biện:06/7/2015; Ngày duyệt đăng: 30/7/2015
Phản biện khoa học: *PGS.TS Nguyễn Văn Dự - Trường Đại học Kỹ thuật Công nghiệp - ĐHTN*

[*] *Tel: 0986 439559, Email: vuvietvu@tnut.edu.vn*