Computers, Materials & Continua CMC, vol.64, no.1, pp.207-216, 2020

Analysis of Semi-Supervised Text Clustering Algorithm on Marine Data

Yu Jiang[1, 2], Dengwen Yu[1], Mingzhao Zhao[1, 2], Hongtao Bai[1, 2], Chong Wang[1, 2, 3] and Lili He[1, 2, *]

Abstract: Semi-supervised clustering improves learning performance as long as it uses a small number of labeled samples to assist un-tagged samples for learning. This paper implements and compares unsupervised and semi-supervised clustering analysis of BOA-Argo ocean text data. Unsupervised K-Means and Affinity Propagation (AP) are two classical clustering algorithms. The Election-AP algorithm is proposed to handle the final cluster number in AP clustering as it has proved to be difficult to control in a suitable range. Semi-supervised samples thermocline data in the BOA-Argo dataset according to the thermocline standard definition, and use this data for semi-supervised cluster analysis. Several semi-supervised clustering algorithms were chosen for comparison of learning performance: Constrained-K-Means, Seeded-K-Means, SAP (Semi-supervised Affinity Propagation), LSAP (Loose Seed AP) and CSAP (Compact Seed AP). In order to adapt the single label, this paper improves the above algorithms to SCKM (improved Constrained-K-Means), SSKM (improved Seeded-K-Means), and SSAP (improved Semi-supervised Affinity Propagationg) to perform semi-supervised clustering analysis on the data. A DSAP (Double Seed AP) semi-supervised clustering algorithm based on compact seeds is proposed as the experimental data shows that DSAP has a better clustering effect. The unsupervised and semi-supervised clustering results are used to analyze the potential patterns of marine data.

Keywords: Unsupervised learning, semi-supervised learning, text clustering.

1 Introduction

With the development of science and technology, marine research is getting more and more mature [Malakoff (2003); Santos, Hawkins, Monteiro et al. (1995); Xie, Ren, Pang et al. (2019)]. Many countries have established marine development strategies for the

[1] College of Computer Science and Technology, Jilin University, Changchun, 130012, China.

[2] A Key Laboratory of Symbolic Computation and Knowledge Engineering, Jilin University, Changchun, 130012, China.

[3] Department of Engineering Mechanics, State Marine Technical University of St. Petersburg, St. Petersburg, 190008, Russia.

[*] Corresponding Author: Lili He. Email: helili@jlu.edu.cn.

Received: 22 January 2020; Accepted: 12 February 2020.

CMC.doi:10.32604/cmc.2020.09861 www.techscience.com/journal/cmc

21st century with real-time monitoring and analysis of marine data. Typically, this data comes in large quantity and has such characteristics as being diverse and inaccurate. Therefore, its processing and analysis will not keep pace with the development of observational technology, and thus limits the exploration of its maximum value. A more comprehensive data mining technology is urgently needed to analyze the ocean temperature, salinity, hydrology and other marine data, in which we can find potential and useful information.

Clustering is an important branch of data mining [Sturn, Quackenbush and Trajanoski (2002); Kaufman and Rousseeuw (2009)]. It is the process of dividing a dataset into multiple datasets of different categories, each of which is called a cluster. The purpose of clustering is to put similar data in the same cluster and make each cluster differ from others. Clustering is an unsupervised learning algorithm, which focuses on the unlabeled sample without prior information. It can divide similar data samples with some similar characteristic attributes into a single dataset. Supervised learning is a kind of data mining algorithm that trains the labeled samples with prior information to obtain the classifier first and then anticipate the categories of unlabeled samples. Unlike unsupervised learning, supervised learning requires a few labeled samples to train, and will have a classifier with better performance when training with large-scale, high-quality samples. In contrast, unsupervised learning obtains different categories after training with unlabeled samples. It usually gets different results, depending on which algorithm is being used. Therefore, the evaluation standard will be different.

However, in some practical problems, supervised learning and unsupervised learning [Zhu and Goldberg (2009); Jordan and Rumelhart (1992)] will be partially limited. This is because the unlabeled samples account for a large proportion of the dataset, whereas the labeled samples are limited and small. It will make the supervised learning model, which should be fruitful and mature, not reach the expected effect. Meanwhile, traditional unsupervised learning cannot effectively use the labeled sample information to assist in the classification of unlabeled samples because it can only classify samples with a target function. Semi-supervised learning has emerged in response to such problems. Semi-supervised learning is the use of a small number of labeled data samples to assist with a large number of unlabeled samples for clustering analysis. Compared with unsupervised and supervised learning algorithms, the advantage of semi-supervised learning is the ability to use both labeled and unlabeled samples, and to make full use of various sample information to improve learning performance. In this paper, the study focuses on the marine text data, analyzes the data with clustering and semi-supervised clustering, and tries to evaluate and interpret the results. Based on the two classical clustering algorithms K-Means [Kanungo, Mount, Netanyahu et al. (2002)] and AP Clustering algorithms, we implement a variety of semi-supervised clustering algorithms. We compare the algorithms with relevant algorithms and try to figure out some potential rules from the results. At the same time, we provide help to the applications of semi-supervised clustering algorithm on marine text data analysis.

The structure of the article is organized as follows. The second section overviews the relevant research of the semi-supervised clustering algorithm, as well as the Argo data set. The theory of semi-supervised clustering algorithm is introduced in the third section.

The experimental results are shown in the fourth section, and the experimental results are briefly analyzed. The final part proposes relevant conclusions.

2 Related work

K-Means clustering is a method commonly used to automatically partition a data set into k groups. It proceeds by selecting k initial cluster centers and then iteratively refining them. In 2007, Arthur et al. [Arthur and Vassilvitskii (2007)] proposed an algorithm augmenting K-Means with a very simple, randomized seeding technique which improves both the speed and the accuracy of K-Means. In 2001, Bradley et al. [Bradley, Bennett and Demiriz (2000)] proposed explicitly adding k constraints to the underlying clustering optimization problem, requiring that each cluster have at least a minimum number of points in it. In 2002, Basu et al. [Basu, Banerjee and Mooney (2002)] explored the use of labeled data to generate initial seed clusters, as well as the use of constraints generated from labeled data to guide the clustering process. It introduces two semi-supervised variants of K-Means clustering that can be viewed as instances of the EM (Expectation Maximization) algorithm, where labeled data provides prior information about the conditional distributions of hidden category labels. In 2017, Keriven et al. [Keriven, Tremblay, Traonmilin et al. (2017)] proposed a compressive version of K-Means (CKM), which estimates cluster centers from a sketch. In 2019, Liu et al. [Liu, Wang, Zhai et al. (2019)] proposed that algorithm integrates imputation and clustering into a unified learning procedure, which achieves superior performance. And the improvement becomes more significant with increasing missing ratio, verifying the effectiveness and advantages of the proposed joint imputation and clustering.

In 2007, Frey et al. [Frey and Dueck (2007)] used affinity propagation to cluster images of faces, detect genes in microarray data, identify representative sentences in this manuscript, and identify cities that are efficiently accessed by airline travel. Affinity propagation found clusters had much fewer errors than other methods, and it did so in less than one-hundredth the amount of time. In 2007, Wang et al. [Wang, Li, Zhang et al. (2007)] proposed semi-supervised affinity propagation, where cluster validity indices are embedded into iteration process of the algorithm to supervise and guide its running to an optimal clustering solution. The experimental results showed that the algorithm gives accurate clustering results for data sets with compact and loose cluster structures. In 2012, Shang et al. [Shang, Jiao, Shi et al. (2012)] proposed a novel Fast Affinity Propagation clustering approach (FAP). FAP simultaneously considers both local and global structure information contained in datasets, and is a high-quality multilevel graph partitioning method that can implement both vector-based and graph-based clustering. In 2016, Jia et al. [Jia, Yu, Wu et al. (2016)] proposed an improved cuckoo search (ICS) technique to solve the AP model. The ICS algorithm utilizes quaternions to represent individuals that are to be optimized. The variable step length of Lévy flights and a method of discovering probability are also proposed. The proposed adaptive AP based on ICS is utilized (or tested) to identify four standard test datasets, such as face images and handwritten digits.

3 Experiments

3.1 K-Means clustering

Now set the K-Means cluster number K to 3. The clustering results are shown in Figs. 1 to 4.

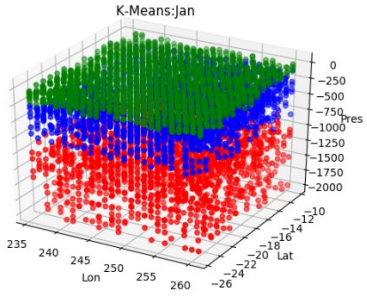

Figure 1: January

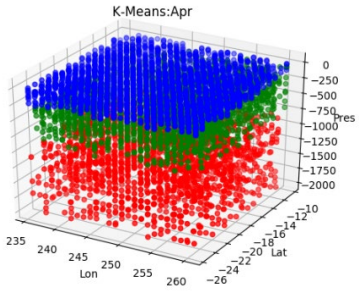

Figure 2: April

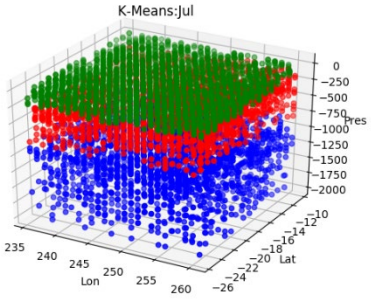

Figure 3: July

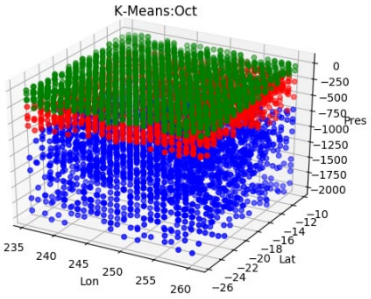

Figure 4: October

lon represents the longitude, *lat* represents the latitude of the southern hemisphere, and *pres* represents the depth (Unit: *dbar*). Because it is below sea level, it takes a negative value. As we can see from Figs. 1 to 4, the marine data points are clearly divided into three layers according to depth: shallow layer 0 *dbar* ~100 *dbar*, middle layer 100 *dbar* ~500 *dbar*, and deep layer greater than 500 *dbar*. This indicates the temperature and salinity influence the vertical layer of marine data. We use the silhouette coefficient to evaluate the K-Means where K takes 3 or 5. Silhouette coefficient shows the degree of closeness in clusters and separation between clusters. Two different *K* comparisons are shown as Fig. 5.

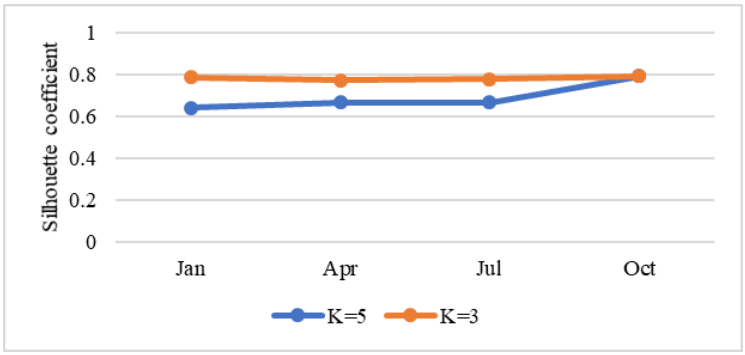

Figure 5: K-Means cluster evaluation of different K

From Fig. 5, we can see that the silhouette coefficient is generally higher when K takes 3 than K takes 5, which indicates that the clustering effect is better when K takes 3 than 5. Therefore, different value of K will affect the clustering performance.

3.2 Election-AP clustering analysis

We use Election-AP to control the number of clusters, which is set to 5, and try to control the number of clusters within 5. Clustering results are shown in Figs. 6 to 9.

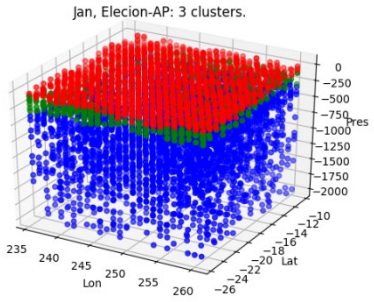

Figure 6: January

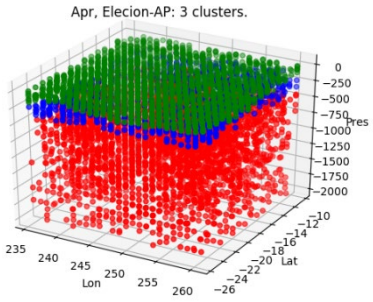

Figure 7: April

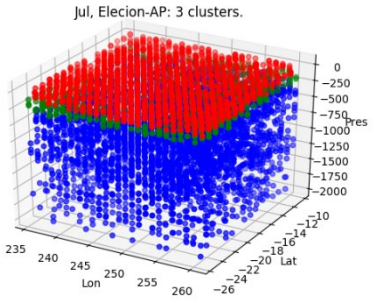

Figure 8: July

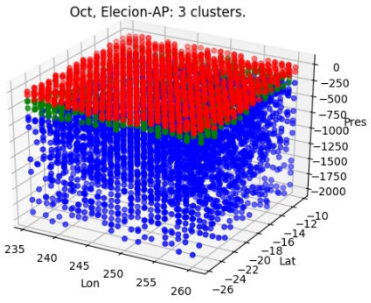

Figure 9: October

In the result figures of Election-AP clustering, we can observe that, under the influence of temperature and salinity, the marine data samples are divided into 3 layers: shallow, middle and deep. Compared with the clustering effect of K-Means algorithm, where

K takes 3, the range of the middle layer becomes narrower. The clustering results of AP and Election-AP are evaluated by silhouette coefficient, which are shown in Fig. 10.

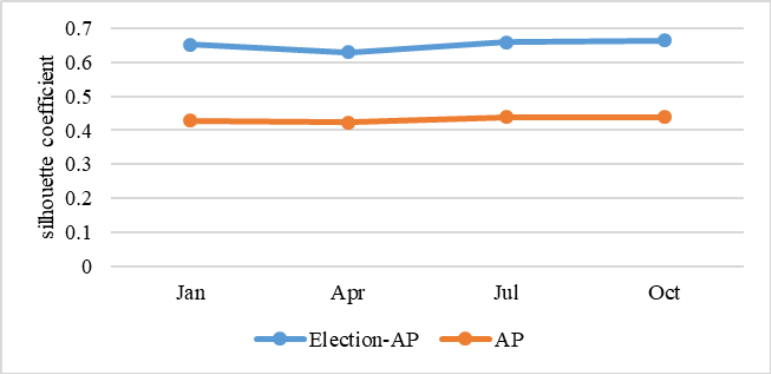

Figure 10: Comparison between Election-AP and AP

From Fig. 10, the performance of AP and Election-AP is relatively stable. The performance of Election-AP clustering is better because it can control the number of clusters in a proper range.

3.3 Semi-supervised clustering analysis

In this section, we select data from January, April, July and October 2016 with a small amount of thermocline labels for semi-supervised clustering and comparison in algorithms. We compare the following algorithms: SCKM, SSKM, SSAP, LSAP, CSAP, DSAP. The dataset size is 5,800 and the size of thermocline samples is 800. We analyze the data of January 2016 with semi-supervised clustering. The results are shown in Figs. 11 to 16, where the blue part is the cluster corresponding to the thermocline.

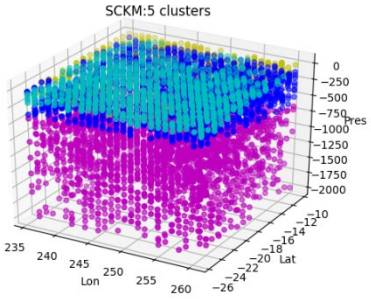

Figure 11: SCKM clustering

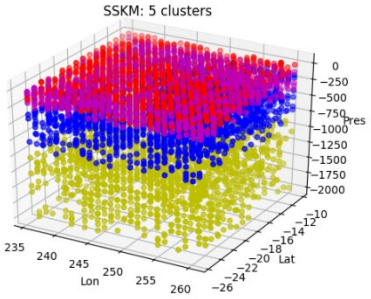

Figure 12: SSKM clustering

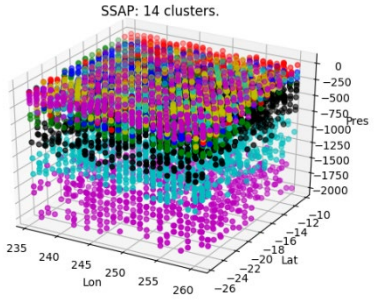

Figure 13: SSAP clustering

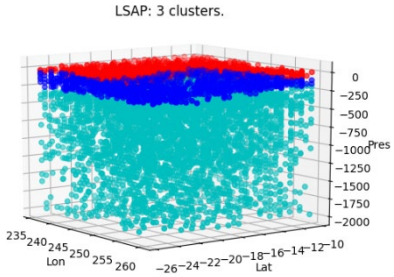

Figure 14: LSAP clustering

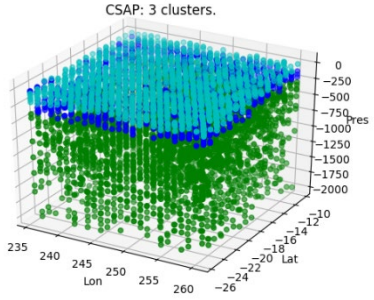

Figure 15: CSAP clustering

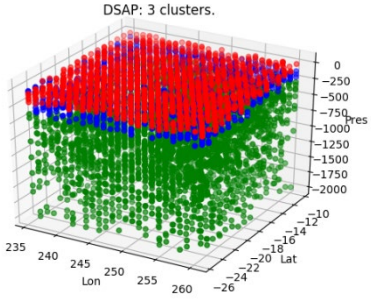

Figure 16: DSAP clustering

From Figs. 11 to 16, we can observe that the data of thermocline is distributed between 100 *dbar* and 200 *dbar* .

Then, we use the SCKM, SSKM, SSAP, LSAP, and CSAP algorithms to do the semi-supervised clustering on the data of April, July and October 2016. We use F to evaluate the results which are shown in Fig. 17.

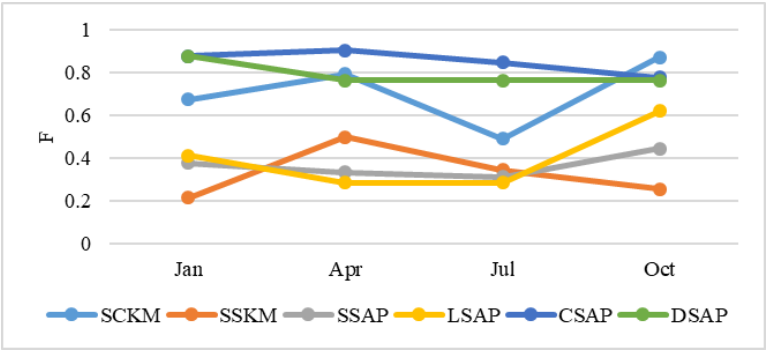

Figure 17: Comparison of F of each semi-supervised algorithm

From the comparison in Fig. 17, we can observe that CSAP and DSAP show a good semi-supervised clustering performance and are stable. SCKM and LSAP are fluctuant, while SSAP and SSKM are stable but their results are not very good. Depending on whether the thermocline samples are reclassified, the semi-supervised clustering

algorithm mentioned above can be divided into two categories: the algorithms that directly allocate the thermocline samples into the same cluster, including SCKM, SCAP, DSAP, and the algorithms that allocate the thermocline samples into different clusters depending on their similarity, including SSKM, SSAP and LSAP. SSKM, SSAP and LSAP may allocate the mislabeled samples to the clusters where they should have been allocated. However, this reduces the recall rate of the thermocline samples. Therefore, SSKM, SSAP and LSAP don't have a good performance when measuring F.

Next, we do the comparison experiment with different sizes of thermocline samples of January 2016 data. In these thermocline samples, randomly select 200, 400, 600 or 800 as the labeled samples, while the rest of these samples are viewed as unlabeled samples. Fig. 18 shows the comparison of F in algorithms that have different size of labeled thermocline samples.

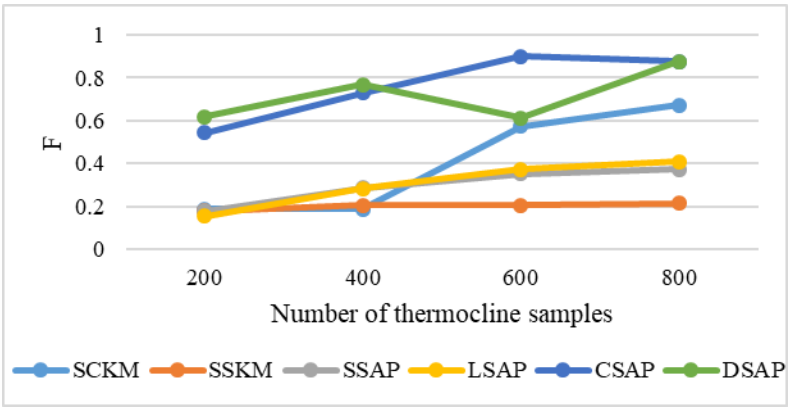

Figure 18: Comparison of F in algorithms that have different size of labeled thermocline samples

From Fig. 18, we can observe that with the increase of the size of labeled thermocline samples, the performance for each semi-supervised algorithm has been improved, of which the improvement of SCKM is notable. When the labeled sample size is relatively small, DSAP and CSAP are better than other semi-supervised algorithms.

4 Conclusions

In this paper, we mainly compare unsupervised clustering algorithm with semi-supervised clustering algorithm through cluster BOA-Argo marine text data, and evaluate the results. In order to reduce the number of clusters of AP clustering and make the results easy to understand, we propose Election-AP based on the idea of re-clustering the center of clusters and achieve good results on the dataset. In order to adapt the traditional semi-supervised clustering algorithm to the single-label (thermocline) cluster data, we optimize the traditional algorithms Constrained-K-Means, Seeded-K-Means and SAP, turn them into SCKM, SSKM and SSAP semi-supervised clustering algorithms, and propose the DSAP algorithm, which has a good performance on this dataset. Through the analysis of unsupervised algorithms and semi-supervised algorithms, we have obtained some laws of ocean data and the thermocline: the marine data can be classified into

different layers in the vertical depth under the influence of temperature and salinity. We also determine the depth range of the thermocline.

Funding Statement: This work was supported in part by the National Natural Science Foundation of China (51679105, 61872160, 51809112); "Thirteenth Five Plan" Science and Technology Project of Education Department, Jilin Province (JJKH20200990KJ).

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

Arthur, D.; Vassilvitskii, S. (2007): K-means++: the advantages of careful seeding. *SODA '07: Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, pp. 1027-1035.

Basu, S.; Banerjee, A.; Mooney, R. (2002): Semi-supervised clustering by seeding. *Proceedings of 19th International Conference on Machine Learning*, pp. 27-34.

Bradley, P. S.; Bennett, K. P.; Demiriz, A. (2000): Constrained k-means clustering. *Microsoft Research, Redmond*, vol. 2000, no. 65, pp. 1-9.

Frey, B. J.; Dueck, D. (2007): Clustering by passing messages between data points. *Science*, vol. 315, no. 5814, pp. 972-976.

Jia, B.; Yu, B.; Wu, Q.; Wei, C.; Law, R. (2016): Adaptive affinity propagation method based on improved cuckoo search. *Knowledge-Based Systems*, vol. 111, pp. 27-35.

Jordan, M. I.; Rumelhart, D. E. (1992): Forward models: supervised learning with a distal teacher. *Cognitive Science*, vol. 16, no. 3, pp. 307-354.

Kanungo, T.; Mount, D. M.; Netanyahu, N. S.; Piatko, C. D.; Silverman, R. et al. (2002): An efficient k-means clustering algorithm: analysis and implementation. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, vol. 24, no. 7, pp. 881-892.

Kaufman, L.; Rousseeuw, P. J. (2009): Finding Groups in Data: An Introduction to Cluster Analysis. John Wiley & Sons.

Keriven, N.; Tremblay, N.; Traonmilin, Y.; Gribonval, R. (2017): Compressive k-means. *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 6369-6373

Liu, Y.; Wang, X.; Zhai, Z.; Chen, R.; Zhang, B. et al. (2019): Timely daily activity recognition from headmost sensor events. *ISA Transactions*, vol. 94, pp. 379-390.

Malakoff, D. (2003): Scientists counting on census to reveal marine biodiversity. *Science*, vol. 302, no. 5646, pp. 773-773.

Santos, R. S.; Hawkins, S.; Monteiro, L. R.; Alves, M.; Isidro, E. J. (1995): Marine research, resources and conservation in the Azores. *Aquatic Conservation: Marine and Freshwater Ecosystems*, vol. 5, no. 4, pp. 311-354.

Shang, F.; Jiao, L. C.; Shi, J.; Wang, F.; Gong, M. (2012): Fast affinity propagation clustering: a multilevel approach. *Pattern Recognition*, vol. 45, no. 1, pp. 474-486.

Sturn, A.; Quackenbush, J.; Trajanoski, Z. (2002): Genesis: cluster analysis of microarray data. *Bioinformatics*, vol. 18, no. 1, pp. 207-208.

Wagstaff, K.; Cardie, C.; Rogers, S.; Schrödl, S. (2001): Constrained k-means clustering with background knowledge. *Eighteenth International Conference on Machine Learning. Morgan Kaufmann Publishers Incorporated*, vol. 1, pp. 577-584.

Wang, K. J.; Li, J.; Zhang, J. Y.; Tu, C. Y. (2007): Semi-supervised affinity propagation clustering. *Computer Engineering*, vol. 33, no. 23, pp. 197-198.

Xiao, Y.; Yu, J. (2008): Semi-supervised clustering based on affinity propagation algorithm. *Journal of Software*, vol. 19, no. 11, pp. 2803-2813.

Xie, X.; Ren, J.; Pang, X.; Lei, C.; Chen, H. (2019): Stratigraphic architectures and associated unconformities of Pearl River Mouth basin during rifting and lithospheric breakup of the South China Sea. *Marine Geophysical Research*, vol. 40, no. 2, pp. 129-144.

Zhu, X.; Goldberg, A. B. (2009): Introduction to semi-supervised learning. *Synthesis Lectures on Artificial Intelligence and Machine Learning*, vol. 3, no. 1, pp. 1-130.