Red Tide Forecast with Semi-supervised clustering

Shun Yao Wu^{a,b}, Fengjing Shao^{a,b,*}, Ying Wang^b, Rencheng Sun^b, Yi Sui^b

^aCollege of Automation Engineering, Qingdao University, Qingdao 266071, China

^bCollege of Information Engineering, Qingdao University, Qingdao 266071, China

*Corresponding Author: sfj@qdu.edu.cn

Abstract

The distribution of red tide monitoring data is obviously imbalanced, which usually causes "uniform effect" problem and has negative influence on red tide prediction accuracy. In order to alleviate this problem, this paper focuses on semi-supervised learning, and proposes an improved method for semi-supervised clustering based on metric learning with pairwise constraints. The main idea is to establish harsher conditions for learning cannot-link constraints, so as to increase distances among clusters. The effectiveness of the proposed method is verified on some balanced and unbalanced UCI datasets. Furthermore, we extend our approach and apply it to red tide forecast, and get excellent performance on real dataset, which is extracted from red tide monitoring data during 2003 to 2009. In addition, we find that the concentration of chlorophyll-a, pH value and dissolved oxygen are key factors to red tide forecast.

Keywords: Red tide forecast, Semi-supervised clustering, Metric learning, Pairwise constraints.

1. Introduction

The red tide is a kind of marine disasters, which means that some tiny phytoplankton, protozoa or bacteria suddenly proliferate and aggregate under certain environmental conditions and cause water discoloration during a period of time⁽¹⁾. In recent years, the red tide occurs more and more frequent, and the magnitude and extent of damage becomes more and more serious, which has become a big threaten to the marine ecological balance, fishery resources, marine aquaculture and human healthy. How to accurately predict the occurrence of red tide as soon as possible is a problem to be solved urgently. Thus, it attracted more and more researchers. Generally, they treat red tide forecast as a

two-class classification, and use the composite index⁽²⁾, neural networks⁽³⁾, clustering analysis^(4,5) to determine whether the red tide will occur, or evaluate the probability of occurrence.

However, most of the red tide monitoring data in the ocean database is normal, and the records about occurrence of red tide are still less. For example, in recent ten years, the normal records in the water quality table of the ocean database are more than four thousands, while the records about the occurrence of red tide are only about one hundred. The main reason is that red tide generally occurs between June and August. To our knowledge, all the aforementioned approaches have not paid attention to the imbalance of red tide data. As stated in literatures^(6, 7), it probably causes "uniform effect" problem (the sizes of produced clusters are relatively uniform) when the input data have varied class sizes. Take Figure 1 as an example, red data points is obviously more than blue ones. If we employ k -means algorithm to partition them randomly select initial point, two initial cluster representatives are probably all red, which may make some red points grouping with the nearby blue ones. Therefore, Liang et al⁽⁷⁾ proposed a multi-center clustering algorithm, and discussed with the fuzzy k -means clustering.

Different with Liang's unsupervised strategy, this paper will discuss the problem with the semi-supervised learning mechanism. As shown in Figure 1, if we enlarge the weight of the x axis as much as possible and make sure the initial cluster representatives are not too bad, it probably deserves a satisfactory performance. Therefore, in order to evaluate the attribute weights to scale the dataset, this paper adopts metric learning^(8,9) with prior knowledge in the form of pairwise constraints. We improve Xing's method⁽⁸⁾ by establishing harsher conditions for learning cannot-link constraints, so as to increase the distances among clusters. The effectiveness of the proposed strategy is verified on

some imbalanced and balanced UCI datasets. Furthermore, we extend our approach and apply it to red tide forecast. As red tide forecast is a two-class classification, we derive more pairwise constraints based on the property of cannot-link constraints, which improves the performance of metric learning; on the other hand, we adopt hard-constraint^(10,15) strategy to lead pairwise constraints into clustering process, so as to improve the clustering quality. We get excellent performance on the real dataset, which is extracted from the red tide monitoring data from 2003 to 2009. In addition, we find that the concentration of chlorophyll-a, pH value and dissolved oxygen are key factors to red tide forecast.

Fig 1. Toy Dataset with varied class sizes.

The remaining chapters are organized as follows. In the next section, we elaborate semi-supervised clustering algorithm based on metric learning and verifies the effectiveness of the improved method on UCI datasets. Next in Section 3, we introduce our approach for red tide forecast based on semi supervised clustering; finally, Section 4 concludes the paper.

2. Semi-supervised Clustering based on Metric Learning

Generally, the traditional clustering divides the objects on the basis of objective indicators (distances/similarities), which does not require training data and has high efficiency. However, its performance is poor, the result is difficult to understand, and it cannot integrate users' requirements. In order to solve these problems, researchers propose semi-supervised clustering, which incorporates a small amount of prior knowledge, so as to guide clustering and get better performance⁽⁸⁻¹⁰⁾. The learning mechanism with a small amount of labeled data and a large number of

unlabeled data is suitable for real application requirements, such as Web Mining, thus has gradually been paid more and more attention to.

2.1 Metric Learning

Metric learning is an important semi-supervised clustering strategy, which constructs optimization to satisfy the prior knowledge as much as possible and obtain new metric via the solution. In 2002, Xing et al.⁽⁸⁾ made important contribution to metric learning. They proposed an effective strategy incorporating pairwise constraints and applied it to clustering analysis. Compared to pointwise constraints, pairwise constraints is more suitable for clustering analysis. The reason is that clustering focuses on similarities between objects rather than the specific categories of objects; on the other hand, the assignment for pairwise constraints is much easier than pointwise constraints in applications.

Definition: must-link constraints set S and cannot-link constraints set D are defined as follows:

- if objects x and y are in the same cluster, $(x, y) \in S$;
- if objects x and y are in different clusters, $(x, y) \in D$.

As shown in formula (1), Xing et al. constructed the optimization with pairwise constraints, which ensures that pairs of points in S have small squared distance between them while the constraint is to make sure that sum of distances between pairs of points in D is not too small. It's because that Xing et al. aimed to enhance tightness between pairs of points in S , which is also usually adopted in lots of work^(11,12). However, our main idea is to increase the distances between clusters, thus Xing's solution for learning cannot-link constraints is no longer suitable. We should pay enough attention to cannot-link constraints, especially when distances between pairs of points in D are larger.

$$\begin{aligned} \min_A \quad & \sum_{(x,y) \in S} \|x - y\|_A^2 \\ \text{s.t.} \quad & \sum_{(x,y) \in D} \|x - y\|_A \geq 1 \end{aligned} \quad (1)$$

where $\|x - y\|_A$ is the Mahalanobis distance, $\|x - y\|_A =$

$$\sqrt{(x - y)^T A (x - y)}.$$

In fact, differently treating the instance level knowledge has been a hotspot of supervised learning and semi-supervised learning. For instance, Karasuyama et al.⁽¹³⁾ identified better classification hyperplane through exploring the weight of pointwise constraints; Basu et al.⁽⁹⁾ set

violations of pairwise constraints based on distances between pairs of points. Therefore, this paper revises the constraint to give harsher lower bounds when distances between points of cannot-link constraints are larger. Firstly, we use the Euclidean distance $d(x, y) = \|x - y\|_2$ to calculate distances of all pairs of points. Denote the minimum distance as $d_{\min}$. Then, we set corresponding constraints for

each cannot-link constraint, viz. $\|x - y\|_2 \geq \frac{d(x, y)}{d_{\min}}$.

2.2 Algorithm Procedure

When A is diagonal, the Mahalanobis distance can be transformed into parameterized Euclidean distance. And, we can transform metric learning into learning attribute weights.

Given dataset X , the number of clustering k , must-link constraints set S and cannot-link constraints set D , the semi-supervised clustering algorithm based on metric learning is as follows:

Step 1 Construct and solve the optimization as formula (1), so as to obtain attribute weights w ;

Step 2 Scale dataset X according to the attribute weights w , $X' = X \cdot w$;

Step 3 Adopt Basu's strategy⁽⁹⁾, use weighted farthest first selection method with pairwise constraints to initialize cluster representatives;

Step 4 Partition scaled dataset X' through k -means clustering algorithm.

2.3 Experimental Results

We compare k -means, Xing's method and our approach on some UCI datasets¹ to verify the effective of the proposed method. For robustness, each result in the experiment is averaged from 10 trials. In addition, we choose two classic external indexes, Purity and NMI, to evaluate clustering quality.

Table 1. Dataset characteristics.

Dataset	n	d	k
glass	214	9	7
diabetes	768	8	2
ionosphere	351	33	2
pageblocks	5473	10	5
wine	178	13	3

vowel	990	10	11
-------	-----	----	----

As shown in Figure 2, we randomly select 5%² must-link constraints and 5% cannot-link constraints for Xing's method and our approach, and the performance of our approach on six datasets is superior to k -means algorithm and Xing's method, especially under NMI index. On diabetes and vowel datasets, Xing's method is worse than k -means algorithm under Purity index. On the one hand, it is because that the prior knowledge is little, especially Xing's method mainly focuses on must-link constraints; on the other hand, inconsistency between must-link and cannot-link constraints may reduce clustering quality⁽¹⁴⁾. Even with such little knowledge, our approach still improves clustering quality through effectively learning cannot-link constraints.

(a)

(b)

Fig 2. Clustering accuracy on UCI datasets.

In addition, this paper also verifies the scalability of

¹ <http://archive.ics.uci.edu/ml/>

² 5% means 0.05n

our approach on vowel dataset. As shown in Figure 3, with the growth of number of pairwise constraints, the proposed method can achieve a more stable growth, and can converge with little knowledge.

Fig 3. Clustering accuracy versus pairwise constraints on vowel datasets.

3. Red Tide Forecast based on Semi-supervised clustering

Red tide forecast is a two-class (red tide class and normal class) classification problem. Firstly, we should make sure that semi-supervised clustering algorithm is able to accomplish the task of identifying category. It can be solved though providing some pointwise constraints. The categories of clustering result can be identified according to number of red tide instances in clusters. In addition, derive more pairwise constraints based on the property of cannot-link constraints, so as to improve the effect of metric learning.

3.1 Red Tide Forecast

In order to satisfy requirements of red tide forecast, we extend the algorithm proposed in section 2, so as to identify the categories of clusters.

Clustering analysis aims to calculate the similarity between objects without concerning the categories of objects. Therefore, after obtaining clustering results, we count number of red tide instances in each cluster, then, treat the cluster with more samples of red tide as red tide class.

As red tide prediction is a two-class classification, pairwise constraints have some interesting properties. For instance, if (x, y) and (x, z) belong cannot-link constraints set D , (y, z) should be a must-link constraint; if (x, y) belong cannot-link constraints set D while (x, z) belong must-link constraints set S , (y, z) should be a cannot-link constraint. On basis of properties of pairwise constraints, we derive more prior knowledge, which effectively improve the performance of initialization and clustering quality.

Moreover, metric learning is independent of clustering process which results that some pairwise constraints are still not satisfied after partitioning. Thus, we adopt hard constraint strategy^(10,15), and modify cluster label of instances during clustering process so as to satisfy pairwise constraints as much as possible.

3.2 Experimental Analysis

We generate red tide dataset through extracting data from the ocean database from 2003 to 2009. Firstly we extract the water temperature, PH value, Vibrio total number and other features from red tide organisms table, hydrometeorological table and water quality table. Secondly, we transform some nominal features, such as date, time, to numeric feature. The missing values are replaced with average value of its corresponding feature. Finally, we get a red tide dataset consist of 615 instances, including 458 normal instances and 157 red tide instances.

Figure 4 is comparison of Xing's method, the proposed method in Section 2 (Our approach 1) and our approach in section 3 (Our approach 2) on the red tide dataset. We select 30% instances as training data. As shown in Figure 4, Our approach 1 is better than the Xing's method, while the improvement is not obvious. In contrast, Our approach 2 has obvious enhancement with more generated pairwise constraints. The hard constraints strategy also plays an important role, which makes almost the entire pairwise constraints can be satisfied. In addition,

we find that the attribute weights of the concentration of chlorophyll-a concentration, pH value and dissolved oxygen are much larger, which should be important factors for red tide prediction.

Fig 4. Prediction accuracy on red Tide dataset.

4. Conclusions

The distribution of red tide monitoring data is obviously imbalanced, which usually causes "uniform effect" problem and has negative influence on red tide prediction accuracy. Different with unsupervised strategy, this paper focuses on semi-supervised learning and improves metric learning with pairwise constraints. We set harsher constraint conditions according to distances between pairs of points in cannot-link constraints set, so as to enlarge the distance between clusters as much as possible. The effectiveness of the proposed method is verified on UCI datasets. Furthermore, we extend our approach to identify categories of clusters for red tide forecast. Moreover, we generate more prior knowledge on basis of properties of pairwise constraints, and adopt hard constraint strategy to satisfy all the pairwise constraints as much as possible. Finally, the effective of the extended approach is verified on Red tide monitoring data from 2003 to 2009. In addition, we find that the concentration of chlorophyll-a, pH value and dissolved oxygen are key factors for red tide prediction.

Acknowledgment

This work was partially supported by the State Key Program of National Natural Science of China (No. 91130035), the National Natural Science Foundation of

China (No. 60974085), the National Public Benefit Research Foundation (No. 200905030), the National Science Foundation of Shandong Province (ZR2012FZ003) and the Science and Technology Planning Project (No. 12-1-4-4-(8)-jch) of Qingdao. Thanks North China Sea Branch of the State Oceanic Administration for providing the red tide data.

References

- (1) http://en.wikipedia.org/wiki/Red_tide
- (2) J. Feng, H. Wang, and J. Sun: "Present Research Condition and Outlooks of the Red tide Forecast in Bohai Sea", Meteorological Hydrological and Marine Instrument, Vol. 1, No. 1, pp. 002, 204
- (3) S. M. Gu, X. H. Sun, Y. H. Wu, and Z. D. Cui: "An approach to forecast red tide using generalized regression neural network", 8th International Conference on Natural Computation (ICNC), pp. 194-198, 2012
- (4) J. Ye, S. Chen, L. Zhou, and J. Ye: "Forecasting red tide occurring level by use of Euclidean distance in Zhoushan sea area", Marine Environmental Science, Vol. 1, pp. 025, 2010
- (5) Y. Zhang: "State Prediction of Algae Reproduction Based on Learning Algorithm of Fuzzy Nearest Neighbor Clustering", Advanced Materials Research, Vol. 356, pp. 929-932, 2012
- (6) H. Xiong, J. Wu, and J. Chen: "K-means clustering versus validation measures: a data-distribution perspective", IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics, Vol. 3, No. 2, pp. 318-331, 2009
- (7) J. Liang, L. Bai, C. Dang, and F. Cao : "The K-Means-Type Algorithms Versus Imbalanced Data Distributions", IEEE Transactions on Fuzzy Systems, Vol. 20, No. 4, 2012
- (8) E. P. Xing, A. Y. Ng, M.I. Jordan, and S. Russell: "Distance metric learning, with application to clustering with side-information", Advances in neural information processing systems, Vol. 15, pp. 505-512, 2002
- (9) S. Basu, M. Bilenko, and R.J. Mooney : "A probabilistic framework for semi-supervised clustering", Proceedings of the 10th ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 22-25, 2004

- (10) A. Huang, D. Miline, E. Frank, and I. H. Witten: "Clustering documents with active learning using wikipedia", 8th IEEE International Conference on Data Mining, pp. 839-844, 2008
- (11) A. Bar-Hillerl, T. Hertz, N. Shental, and D. Weinshall: "Learning a mahalanobis metric from equivalence constraints", Journal of Machine Learning Research, Vol. 6, No. 1, pp.937, 2006
- (12) X. S. Yin, E. L. Hu, and S. C. Chen: "Discriminative semi-supervised clustering analysis with pairwise constraints", Journal of Software, Vol. 19, No. 11, pp. 2791-2802, 2008
- (13) M. Karasuyama, N. Harada, M. Sugiyama, and I. Takeuchi: "Machine Learning for Signal Processing (MLSP)", IEEE International Workshop on Machine Learning for Signal Processing, pp. 1-6, 2011.
- (14) I. Davidson, K. Wagstaff, and S. Basu: "Measuring constraint-set utility for partitional clustering algorithms", Knowledge Discovery in Databases: PKDD 2006, pp. 115-126, 2006
- (15) K. Wagstaff, C. Cardie, S. Rogers, and S. SchrodL: "Constrained k-means clustering with background knowledge", Proceedings of the 8th International Conference on Machine Learning, pp. 577-584, 2001