A Review of Ensemble Based Classification and Clustering in Machine Learning

Rupali M. Bora[1], Smita N. Chaudhari[2], Shilpa P. Mene[3]
[1,2,3]*Department of Information Technology,*
K. K. Wagh Institute of Engineering Education & Research, Nashik

Abstract-Ensemble methods have been effectively used as a classification scheme. Bagging is a simple and powerful ensemble method which relies on bootstrap sampling over training data for diversity production. In ensemble margin framework, the margin theory is exploited to design better ensemble classifiers. Due to the ability to combine multiple base clusterings into a probably better and more robust clustering, the ensemble clustering technique has been attracting increasing attention in recent years. Despite the significant success, one limitation to most of the existing ensemble classification and clustering methods is that they generally treat all base models equally regardless of their reliability, which makes them vulnerable to low-quality base classifiers and clustering. It remains an open problem how to evaluate the reliability of these models and exploit the local diversity in the ensemble to enhance the consensus performance, especially, in the case when there is no access to data features or specific assumptions on data distribution. In this paper a review of margin-based classifiers and clustering approach based on ensemble-driven cluster uncertainty estimation and local weighting strategy is presented.

Keywords: ensemble-based classification, margin, bagging, local weighting, ensemble clustering

I. INTRODUCTION

Ensemble learning is a powerful learning prototype, which builds a classification model by integrating multiple diversified constituent models. The achievement of ensemble methods arises primarily from the fact that they improve the overall predictive performance. Usually, ensemble methods consist of two phases : the production of multiple classifiers and their combination. An ensemble can be composed of either homogeneous or heterogeneous classifiers. Homogeneous classifiers, which are the most widely used, derive from different executions of the same learning algorithm. Such classification models can be produced for example through the manipulation of the training instances or the input attributes.

Data clustering is a fundamental but a very demanding problem in the field of data mining and machine learning. The purpose of it is to discover the inherent structures of a given dataset and partition the dataset into a certain number of homogeneous clusters. During the past few decades, a large number of clustering algorithms have been developed by exploiting various techniques. No clustering algorithm is capable of dealing with all types of data and cluster shapes. If a data set is given, different clustering algorithms, or even the same algorithm with different initializations or parameters, may lead to different clustering results. However, without prior knowledge, it is extremely difficult to decide which algorithm would be the appropriate one for a given clustering task. Even with the clustering algorithm given, it may still be tricky to find proper parameters for it.

II. MARGIN-BASED CLASSIFICATION

2.1 Bagging

A bagging algorithm, relying on an unsupervised ensemble margin, is then applied to minimize data redundancy and improve classification accuracy. Bagging is one of the most successful ensemble methods. It relies on bootstrap sampling over training data to produce diversity. Diversity is derived from the differences between the training sets of base classifiers. The more differences in these training sets the more diversity can be achieved. Redundancy not only slows down the training task but it can also significantly decrease diversity, thus degrading the performance of bagging, affecting the rarer and most difficult classes. In the training process, each example carries its own piece of contribution about the target. However, the contributions are different from each other. Obviously, the redundant instances' contribution is less significant than the contribution of other unmatched instances. [1]

2.2 Margin-based bagging

The group the training instances into three categories: typical, critical and noisy. Generally, especially in image classification, typical instances occur in larger numbers than critical instances such as class decision boundary instances. Consequently, typical instances are more likely to be similar. Thus, there is potentially more redundancy in these instances. The bagging algorithm is based on training instance ordering and relies on the unsupervised ensemble margin previously defined. This technique selects the most informative training instances based on their

margins. This method orders the training instances according to their margin values; the higher the margin of a training instance, the higher the probability this instance being typical and potentially redundant.[1]
It trains first the group of bootstrapped base classifiers on the complete training set and then removes iteratively a fixed portion of the data points on which the committee most agrees according to the unsupervised ensemble margin previously defined. This amounts to removing typical instances during the learning process. The training of the bagging ensemble takes place again with the resulting reduced training set composed of the lowest margin instances. New margin values are calculated for each remaining training instance. This iterative guided procedure is repeated until reaching a maximum training accuracy providing an optimal ensemble designed with a reduced and more informative training set [1].

2.3 Ensemble Margin For Diversity
The success of ensemble learning is credited to the complementarily of each base classifier, making errors on different instances in the ensemble. This difference is also called diversity. The diversity of an ensemble classifier can be measured at input (training data) level, at structure and parameter level, and at output level (predicted). Many diversity measures have been proposed in literature at output level. These different measures are usually grouped into two types: pair wise and non-pair wise. To measure the gain of an ensemble approach with respect to a single classifier, one should focus on the instances on which the ensemble does increase the performance compared to a single classifier. Therefore, using the ensemble margin to express diversity is worthwhile to investigate. A new non-pairwise diversity measure is proposed emphasizing lower margin instances [1].
The ensemble margin [7] is a fundamental concept in ensemble learning. Several studies have shown that the generalization performance of an ensemble classifier is related to the distribution of its margins on the training data [6, 7]. The margin can provide extra information for improving classification accuracy but how to use it is still not yet fully explored. The decision by an ensemble for each instance is made by voting. This vote is the hypothesis of the ensemble decision function. The ensemble margin can be calculated as a difference between the votes, therefore, it is a distance between the hypothesis. The most popular ensemble margin is given by the difference between the fraction of classifiers voting correctly and incorrectly [7].
This margin can be computed by equation 1, where c1 is the most voted class for sample x and vc1 the number of related votes, c2 is the second most popular class and vc2 the number of corresponding votes. This margin's range is from 0 to +1. It is an alternative to the classical definition of the margin [7] with an appealing property : it does not require the true class label of an instance, i.e. it is unsupervised. Thus, it is potentially more robust to noise as it is not affected by errors occurring on the class label itself.[1]

$$\begin{aligned} margin(x) &= \frac{v_{c_1} - v_{c_2}}{\sum_{c=1}^{L}(v_c)} \\ &= \frac{\max_{c=1,\ldots,L}(v_c) - \max_{c=1,\ldots,L \cap c \neq c_1}(v_c)}{T} \end{aligned} \quad (1)$$

A little margin instances have a major influence in building trustworthy classifiers. The margin paradigm is essential for a new ordering-based mislabeled instance elimination method. The same margin framework is used to derive an ensemble diversity measure. Results show that low margin instances have a major influence on forming an appropriate training set to build reliable ensemble classifiers, leading to a significant increase in both overall and per-class accuracies. In supervised learning, the training set is very essential component of the learning process. The main focus should be on forming an appropriate training set made of the most useful examples and without any mislabeled data. Ordering training instances according to their probability of being mislabeled is a simple and efficient method for noise removal. In this approach, these probabilities rely on the margin values of training instances [1].
The mislabeled instance ordering approach simply relies on ensemble margin's definition as a noise evaluation function. Only the training instances xi are assessed whose attribution and label values are not consistent. An ordering-based mislabeled instance elimination and correction method is used as a correction method, where the original class labels are replaced by the most voted class label. The algorithm consists of the following steps [1]:
1. Constructing an ensemble classifier with training data
2. Computing the margin of each training instance
3. Ordering all the training instances, that have been misclassified, according to their noise evaluation values in descending order
4. Eliminating or correcting the first few mislabeled instances to form a new cleaner training set

5. Evaluating the cleaned training set by classification performance, on a validation set
6. Selecting the best filtered training set

III. LOCALLY WEIGHTED CLUSTERING

Although some efforts have been made to (globally) evaluate and weight the base clustering, yet these methods tend to view each base clustering as an individual and neglect the local diversity of clusters inside the same base clustering. In particular, the uncertainty of each cluster is estimated by considering the cluster labels in the entire ensemble via an entropic criterion.

A novel ensemble-driven cluster validity measure is introduced, and a locally weighted co-association matrix is presented to serve as a summary for the ensemble of diverse clusters. With the local diversity in ensembles exploited, two novel consensus functions are further proposed. Extensive experiments on a variety of real-world datasets demonstrate the superiority of the proposed approach over the state-of-the-art [2].

3.1 Methodology

In the general formulation of ensemble clustering here is no access to the original data features. Without needing access to the data features or relying on specific assumptions about data distribution, the key problem here is how to evaluate the reliability of clusters and weight them accordingly to enhance the accuracy and robustness of the consensus clusterings.

Aiming to address the aforementioned problem, in this paper, a novel ensemble clustering approach is proposed based on ensemble-driven cluster uncertainty estimation and local weighting strategy. The overall process of this approach is illustrated in Fig. 1. It takes the advantage of the ensemble diversity at the cluster-level and integrate the cluster uncertainty and validity into a locally weighted scheme to enhance the consensus performance. A cluster can be viewed as a local region in the corresponding base clustering. Without needing access to the data features, in this paper, the uncertainty of each cluster is estimated with regard to the cluster labels in the entire ensemble based on an entropic criterion. In particular, given a cluster, it is investigated the uncertainty by considering how the objects inside this cluster are grouped in the multiple base clusterings. Based on cluster uncertainty estimation, an ensemble-driven cluster index (ECI) is then presented to measure the reliability of clusters. In this paper, it's argued that the crowd of diverse clusters in the ensemble can provide an effective indication for evaluating each individual cluster [2].

By evaluating and weighting the clusters in the ensemble via the ECI measure, the concept of LWCA matrix is presented, which incorporates local adaptively into the conventional co-association (CA) matrix and serves as a summary for the ensemble of diverse clusters. Finally, to achieve the final clustering result, two novel consensus functions are proposed, termed locally weighted evidence accumulation (LWEA) and locally weighted graph partitioning (LWGP), respectively, with the diversity of clusters exploited and the local weighting strategy incorporated [2].

The main contributions of this paper are as follows.

1) The uncertainty of clusters is proposed by considering the distribution of all cluster labels in the ensemble using an entropic criterion, which requires no access to the original data features and makes no assumptions on the data distribution.

2) An ensemble-driven cluster validity index is presented to evaluate and weight the clusters in the ensemble, which provides an indication of reliability at the cluster-level and plays a crucial role in the local weighting scheme.

3) Two novel consensus functions are proposed to construct the final clustering based on ensemble-driven cluster uncertainty estimation and local weighting strategy.

4) Extensive experiments have been conducted on a variety of real-world datasets, which demonstrate the superiority of the proposed ensemble clustering approach in terms of both clustering quality and efficiency [2].

IV. EXPERIMENTAL RESULTS

The ensemble evaluation function which has been used was overall accuracy on training set; if the accuracy of the hardest class as evaluation function instead, the accuracy of the most difficult class on test set would further increase. Thus, our method not only improves the overall classification accuracy compared to classic bagging, but also significantly increases the accuracy of the hardest class. Schapire's margin [4] is one of the most popular margin formulations. It is calculated as the difference between the votes to the correct class label and the maximal votes to any single incorrect class label. The margin-based bagging method, which relies on an unsupervised ensemble margin, outperforms Schapire's margin-based bagging, which relies on a supervised ensemble margin. The involvement of the unsupervised margin in our bagging algorithm significantly improves the ensemble performance for about half the data sets compared to the use of the classic supervised margin [1]. The execution

time of different ensemble clustering methods is compared with varying data sizes. The experiments are performed on different subsets of a dataset.[2]

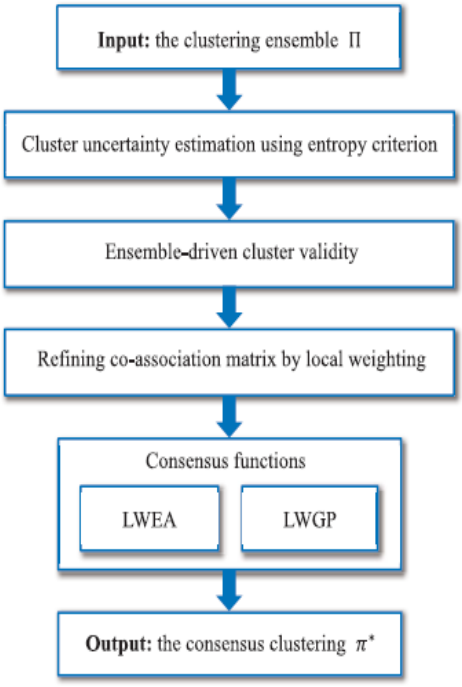

Fig. 1 Clustering approach [2]

V. CONCLUSION

An efficient ensemble design based on an unsupervised version of ensemble margin is reviewed here. A simple bagging method is elaborated, which relies on critical instances, which have low margins, to improve the learning process. This method increases the ensemble performance, particularly in case of difficult or rare classes. Hence, targeting lower margin instances which represent samples closer to class decision boundaries and/or more difficult than higher margin samples demonstrates improved ensemble performance. This strategy reduces data redundancy and increases information significance. Therefore, it designs stronger ensemble classifiers with an increased capability for handling hard or rare classes. An ensemble clustering approach based on ensemble-driven cluster uncertainty estimation and local weighting strategy is also reviewed. The uncertainty of clusters is estimated by considering the cluster labels in the entire ensemble based on an entropic criterion. A new ensemble-driven cluster validity index termed ECI is computed. The ECI measure requires no access to the original data features and makes no assumptions on the data distribution. Then, a local weighting scheme is presented to extend the conventional CA matrix into the LWCA matrix via the ECI measure. With the reliability of clusters investigated and the local diversity in ensembles exploited, two novel consensus functions are proposed, termed LWEA and LWGP, respectively. The authors [D. Huang, C. Wang and J. Lai] have conducted extensive experiments on fifteen real-world datasets. The experimental results have shown the superiority of the approach in both clustering quality and efficiency when compared to the state-of-the-art approaches.

VI. REFERENCES

[1] L. Quo and S. Boukir, "Building an ensemble classifier using ensemble margin. Application to image classification," 2017 IEEE International Conference on Image Processing (ICIP), Beijing, 2017, pp. 4492-4496.
[2] D. Huang, C. Wang and J. Lai, "Locally Weighted Ensemble Clustering," in IEEE Transactions on Cybernetics, vol. 48, no. 5, pp. 1460-1473, May 2018
[3] L. Mason, P.L. Bartlett, and J. Baxter, “Improved generalization through explicit optimization of margins,” Machine Learning, vol. 38, no. 3, pp. 243–255, 2000.
[4] R.E. Schapire, Y. Freund, P. Bartlett, and W.S. Lee, “Boosting the margin: a new explanation for the effectiveness of voting methods,” The Annals of Statistics, vol. 26, no. 5, pp. 1651–1686, 1998.