A survey on ensemble learning Xibin DONG, Zhiwen YU , Wenming CAO, Yifan SHI, QianliMA

Rana Demir 201401003
Zeynep Meriç Aşık 201410026

Abstract

- Complex data problem
  - ✓ Imbalanced
  - ✓ High-dimensional
  - ✓ Noisy
- Aim: Integrating
  - ✓ Data fusion
  - ✓ Data modeling
  - ✓ Data mining
- Better predictive performance by voting schemes

Keywords

- Ensemble Learning
- Supervised Ensemble Classification
- Semi-Supervised Ensemble Classification
- Clustering Ensemble
- Semi-Supervised Clustering Ensemble

Bias-Varience Tradeoff

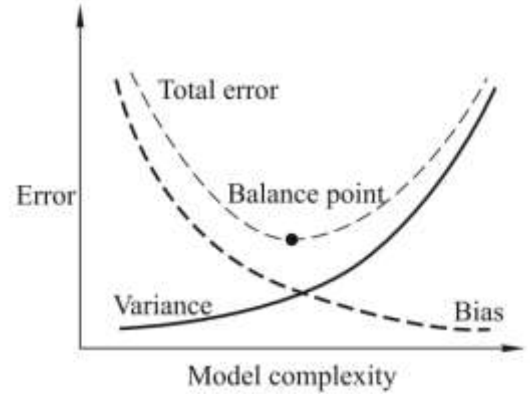


Fig. 1 The relationship between learning curve and model complexity

About Ensemble Learning

- The earliest work: Utilizing the component classifiers trained from dıfferent categories to constitute a composite classification system by Dasarathy and Sheela
- Has exceptionally satisfactory performance in competitions like Kaggle
- Aims to integrate various machine learning algorithms into a unified framework seamlessly

Supervised Ensemble Classificatıon

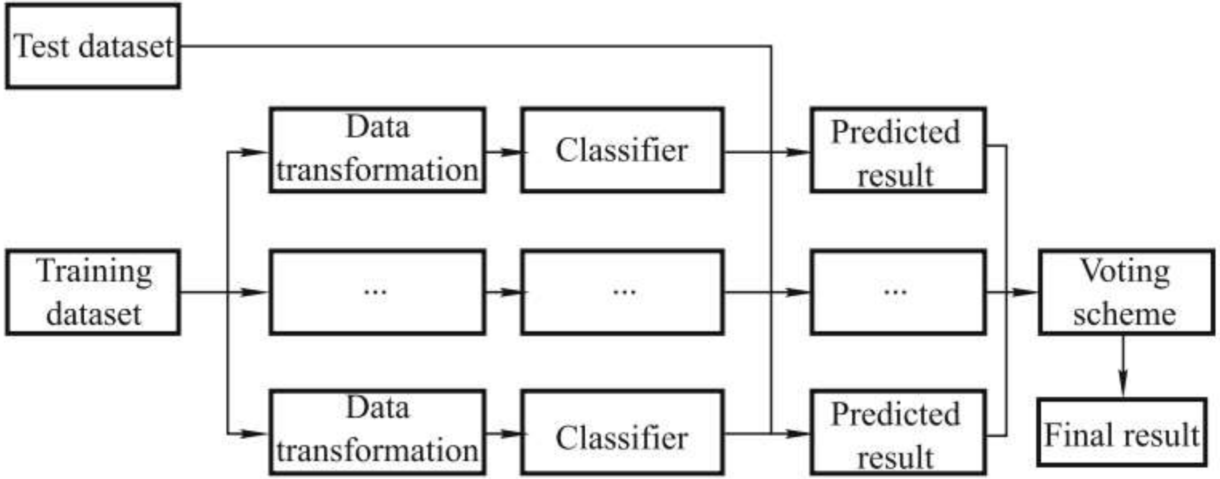


Fig. 4 The framework of ensemble classification

Ensemble Classification Methods

1. Sample Level
2. Feature Level

Some works on

- Feature subset selection,
- Feature extraction,
- Redundancy feature removal etc.

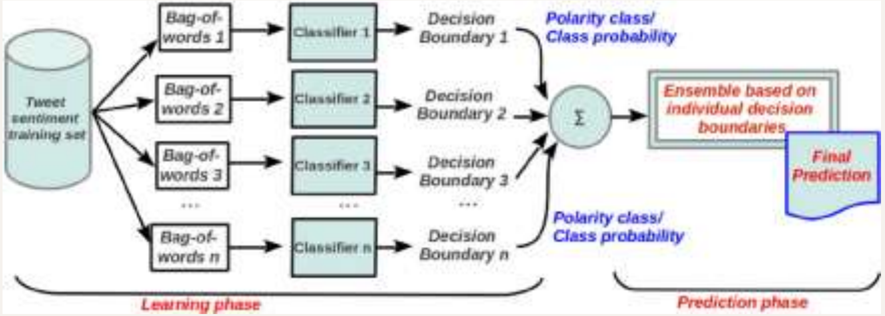
Eliminating basic models

- Some models may not be beneficial to the final result of integratoon.
- Methods:
  - Gasen-b (Zhou and Tang)
  - Diao’s feature selection method

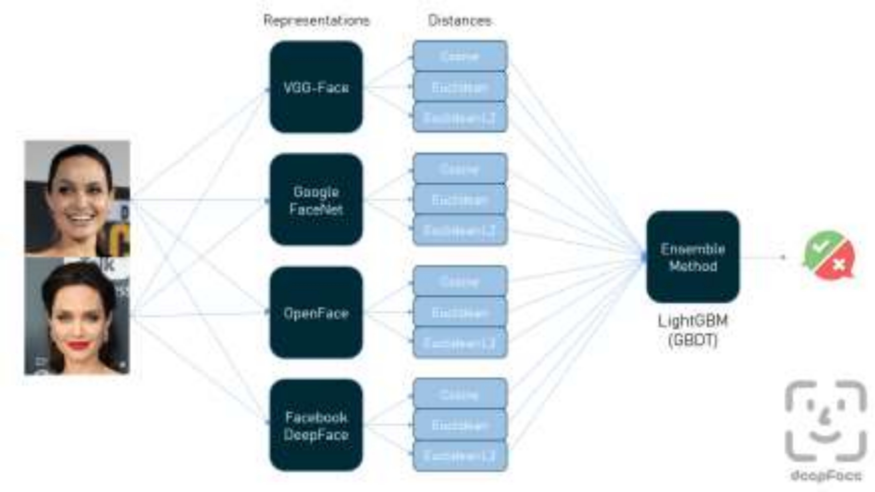
Advantages

- Ensemble classification methods have advantages in terms of
  - ✓ Accuracy,
  - ✓ Stability,
  - ✓ Generalization
- Can be adopted to solve
  - ✓ Multi-instance learning
  - ✓ Multiple-label learning
  - ✓ Imbalance learning

Widely used in



- Biomedical field
- Intelligent transportation area
- Pattern recognition applications
- Social applications



What to improve?

Although most methods mainly consider improving the accuracy of the model at the architecture level of ensemble models, there are quite a few researches on determining the appropriate model size and reducing the complexity of the model to increase the training speed.

Performances of ensemble classifıcation models can be further improved by takıng the interconnection and feedback between different levels.

Semi-supervised Ensemble Classification

Semi-supervised ensemble classification methods focus on expanding the training set and utilizing it.

The mechanism can help capture more accurate underlying data distribution by introducing more informative data thus outperforms other traditional ensemble classification methods in the case where the labeled data is insufficient.

Semi-supervised Ensemble Classification

In the figure below, there is a decision boundary adjusted by semi- supervised method.

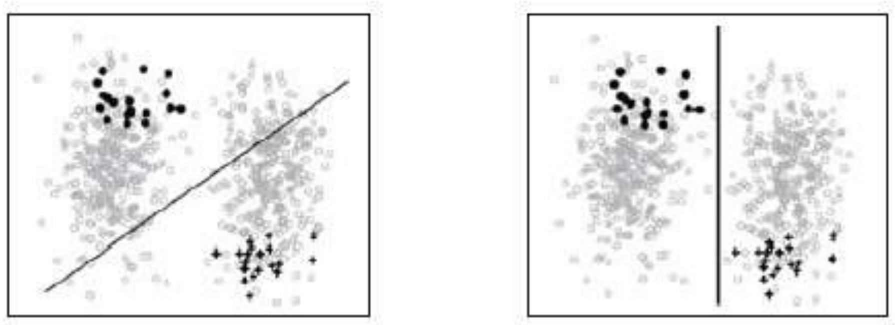


Fig. 8 Decision boundary adjusted by semi-supervised method

Semi-supervised Ensemble Classification

Most of the existing semi-supervised classification models focus on hidden structures, distributional information, dependencies and other characteristics of the data.

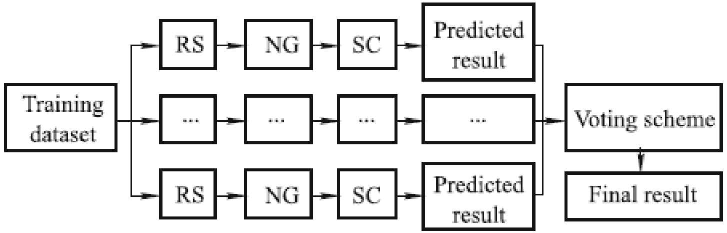


Fig. 9 The framework of semi-supervised ensemble classification in subspace

Some Uses of Semi-supervised Ensemble Classification

- Fault classification in micro-grids system
- Automatic mine detection
- Urban pollutant monitoring
- Spam short message service detection

Clustering Ensemble

Clustering ensemble algorithm works by generating a series of clustering partition using clustering algorithms and combining the partitions together to get the consensus solution.

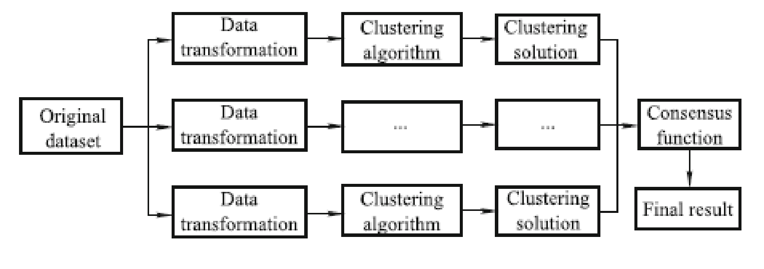


Fig. 11 The framework of clustering ensemble

Clustering Ensemble

The first type of research works focuses on designing new algorithms for the clustering member generation process and the combination process.

This category of researches focuses on investigating delicate clustering ensemble algorithms from different perspectives rather than simply aligning clustering results obtained from traditional algorithms.

Clustering Ensemble

The second type of researches theoretically analyzes clustering ensemble model characteristics such as stability,diversity, and convergence.

The purpose of this type of researches was to improve performances of clustering ensemble algorithms and provide theoretical supports.

Clustering Ensemble

The third type of researches focuses on the selection of clustering results from ensemble models.

Uses of Clustering Ensemble

Clustering ensemble methods have been widely used to solve a diversity of real-world problems on different areas such as data mining, bioinformatics, DNA microarray analysis, gene expression data analysis, image segmentation, significant region detection, biomedical text clustering, speaker recognition, internet security, filtering recommendation etc.

Clustering Ensemble

Concerning traditional clustering ensemble algorithms, there is a lack of theoretical design principles for sample allocations for each clustering member.

In addition, in the case where the prior information is not available, it is still a problem to determine the numbers of clustering members and final clustering members, which needs to be further explored.

Moreover, because the clustering results of ensemble members need to be fused, the time complexity is extremely high, especially when dealing with high-dimensional data. In this case, it is obligatory to develop efficient algorithms to reduce the time complexity of the models.

Semi-supervised Clustering Ensemble

Semi-supervised clustering ensemble can be treated as a technical combination of semi-supervised clustering and ensemble learning. Thus, it allows fusing advantages of both techniques to improve the accuracy and robustness of the model, compared with traditional clustering ensemble methods.

Semi-supervised Clustering Ensemble

Most of the works on semi-supervised clustering ensemble focus on optimizing the generation process and the selection process of clustering members and the algorithm can be summarized with the figure below.

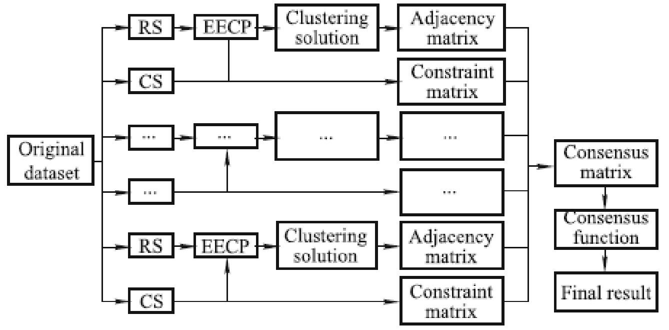


Fig. 12 The framework of incremental semi-supervised clustering ensemble

Semi-supervised Clustering Ensemble

In the field of semi-supervised clustering ensemble, some works focus on optimizing the generation process and the selection process of clustering members.

Some researchers studied the way to integrate clustering members into final prediction via some special voting mechanisms such as collaborative training by using tri-training or adaptive ensemble member weighting process.

Uses of Semi-supervised Clustering Ensemble

- Categorize the web videos
- Social media mining
- Character labeling
- Cancer classification

A Solution to Some Disadvantages

There are some disadvantages like failing to make full use of the constraint information and optimize the constraint selection. However, the semi-supervised mechanism can be refined by introducing unlabeled data from other sources to effectively overcome the shortcomings. Thus, it is worth to explore the refinement of semi-supervised mechanism by other machine learning methods such as transfer learning, active learning and more for clustering ensemble in certain rational formations.

New Direction

Ensemble learning is a relatively mature machine learning issue, compared with recent machine learning hot spots. Previous sections show that ensemble learning is more of a frame ideology, so it is possible for ensemble learning to combine with other machine learning methods smoothly. The works from recent years also prove that ensemble learning can successfully fuse with the other machine learning issues mentioned in the below figure and effectively improved them with its mechanism.

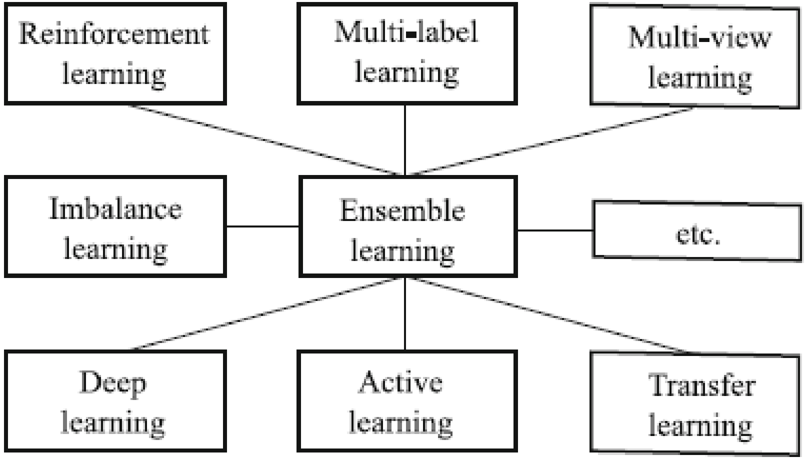


Fig. 13 The combination of ensemble learning with other machine learning issues

New Direction

As for the transfer learning issue, related works used ensemble method to combine outputs of various selective layer based transference conditions of deep learning model. Experiments show it can reduce the effect of negative feature transference on image recognition tasks.

Summary

In this paper, investigations are made about the research progress in various branches of ensemble learning, and categorized methods from different perspectives. Besides, challenges and feasible research directions are introduced for ensemble learning. However, there are still more efforts to make to further improve performances of ensemble models, especially in the case where data contains complex patterns. The expectation is to throw some light on this field by providing some suggestions for future ensemble learning directions.

Thank you for listening