Abstract

One of the main tasks of data mining, clustering is used to group *non-labeled* data to find meaningful patterns. Generally, different models provide predictions with different accuracy rates. Thus, it would be more efficient to develop a number of models using different data subsets, or utilizing differing conditions within the modeling methodology of choice. However, selecting the best model is not necessarily the ideal choice because potentially valuable information may be wasted by discarding the results of less-successful models. This leads to the concept of combining, where outputs (individual predictions) of several models are pooled to make a better decision (collective prediction). Research in the Clustering Combination field has shown that these pooled outputs have more strength, novelty, stability, and flexibility than the results provided by individual algorithms.

Nevertheless, in the social science arena, there is a corresponding research field known as the Wisdom of Crowds, after the book by the same name written by Surowiecki in 2004, simply claiming that the Wisdom of Crowds (WOC) is the phenomenon whereby the decisions made by aggregating the information of groups usually have better results than those made by any single group members. Surowiecki suggested a clear structure for building a wise crowd. Supported by many examples from businesses, economies, societies, and nations, he argued that a wise crowd must satisfy four conditions, namely: diversity, independence, decentralization, and an aggregation mechanism.

This research studies the previous background and related work of cluster ensemble. Furthermore there is a review over the literature of the wisdom of crowds. The purpose of this study is to suggest ways of mapping and using this theory in selecting suitable clusters in the cluster ensemble. Thus, in this study two methods are proposed for combining the two techniques. Firstly, by incorporating the definitions used in the wisdom of crowds, in other words the four conditions of a wised crowd are redefined to be used in the cluster ensemble selection. Then, by using these definitions, the first method will be proposed in which thresholding is used for generating the final result. In this method the primary clustering algorithms with deferent types are considered independently, for which thresholding is needed. In the second method, the two parts of the first approach have been improved. In order to model and evaluate the independence of clustering algorithms, a technique based on algorithm graph code is presented. The degree of obtained independence level from this approach is used as a weight to evaluate diversity in generating the final result. To clarify our claim in this research, the results from the above approaches are compared with basic clustering methods, cluster ensemble methods, and cluster ensemble selection methods, using primarily standard UCI repository data sets. In conclusion section, all proposed methods for future work are also mentioned.

Keywords Cluster ensemble, Wisdom of crowd, Independency of algorithms, Diversity of results, Decentralization of framework