Information Sciences 361–362 (2016) 120–134

Contents lists available at ScienceDirect

Information Sciences

journal homepage: www.elsevier.com/locate/ins

Diversity control for improving the analysis of consensus clustering

Milton Pividori[a,b,*], Georgina Stegmayer[a], Diego H. Milone[a]

[a] Research Institute for Signals, Systems and Computational Intelligence, sinc(i), UNL-CONICET, Department of Informatics, FICH, Universidad Nacional del Litoral, Ciudad Universitaria CC 217, Ruta Nac. No 168, km 472.4, Santa Fe 3000, Argentina
[b] Center of Research and Development of Information Systems Engineering, CIDISI, UTN-CONICET, Department of Information Systems Engineering, FRSF, Universidad Tecnológica Nacional, Lavaisse 610, Santa Fe 3000, Argentina

ARTICLE INFO

Article history:
Received 21 September 2015
Revised 18 March 2016
Accepted 22 April 2016
Available online 28 April 2016

Keywords:
Cluster ensembles
Consensus clustering
Diversity analysis
Diversity control
Ensemble diversity

ABSTRACT

Consensus clustering has emerged as a powerful technique for obtaining better clustering results, where a set of data partitions (ensemble) are generated, which are then combined to obtain a consolidated solution (consensus partition) that outperforms all of the members of the input set. The diversity of ensemble partitions has been found to be a key aspect for obtaining good results, but the conclusions of previous studies are contradictory. Therefore, ensemble diversity analysis is currently an important issue because there are no methods for smoothly changing the diversity of an ensemble, which makes it very difficult to study the impact of ensemble diversity on consensus results. Indeed, ensembles with similar diversity can have very different properties, thereby producing a consensus function with unpredictable behavior. In this study, we propose a novel method for increasing and decreasing the diversity of data partitions in a smooth manner by adjusting a single parameter, thereby achieving fine-grained control of ensemble diversity. The results obtained using well-known data sets indicate that the proposed method is effective for controlling the dissimilarity among ensemble members to obtain a consensus function with smooth behavior. This method is important for facilitating the analysis of the impact of ensemble diversity in consensus clustering.

1. Introduction

Clustering is fundamental for understanding the structure of data [45] and it has been used in a wide range of areas, including engineering, financial, biological science, and medical applications [25,29,34,40,44]. However, the correct choice of a clustering algorithm, or even setting its parameters, requires knowledge of the data set and the data distribution assumed by algorithms, since they can strongly affect the final results obtained [22]. Clustering algorithms have been developed to solve a wide range of different problems, but there is no universal method that can be applied to solve all. Thus, different

* Corresponding author at: Research Institute for Signals, Systems and Computational Intelligence, sinc(i), UNL-CONICET, Department of Informatics, FICH, Universidad Nacional del Litoral, Ciudad Universitaria CC 217, Ruta Nac. No 168, km 472.4, Santa Fe 3000, Argentina. Tel.: +54 342 4575233x 190; fax: +54 3424575224.
E-mail addresses: mpividori@sinc.unl.edu.ar (M. Pividori), gstegmayer@sinc.unl.edu.ar (G. Stegmayer), dmilone@sinc.unl.edu.ar (D.H. Milone).
URL: http://www.sinc.unl.edu.ar (M. Pividori)

http://dx.doi.org/10.1016/j.ins.2016.04.027
0020-0255/

but equally valid solutions can be obtained using various algorithms, which is one reason why clustering is considered to be an ill-posed problem among researchers [24,50].

In the past decade, consensus clustering (or cluster ensembles) has emerged as a powerful approach for mitigating the issues of conventional cluster analysis. First, a set of data partitions is generated, which is called an *ensemble*. Next, a *consensus function* combines the ensemble into a consolidated solution or *consensus partition*, which has greater overall accuracy [10,18,19,21,39,42,54,56]. Given the ill-posed nature of clustering, the accuracy is typically measured by comparing the final solution with a known reference partition, which is generally based on the class labels associated with the data set [20,30,41,46,50]. Although this reference partition might not be the only valid structure for the data, many studies have tried to determine how ensembles should be built, and which characteristics they should have to obtain high accuracy. In particular, among these characteristics, the level of disagreement between ensemble members, which is called the *ensemble diversity*, has been identified as a key factor in the cluster ensemble problem [7,17,20], and various diversity measures have been proposed [2,14,15,33,55].

Many strategies have been used to explore how diversity affects consensus performance [8,20], where they usually aim to generate a set of ensembles with different diversity, before observing the performance of the consensus function. One of the most common approaches involves generating the ensemble members by randomly varying a parameter [8,15,17,27], which can be the clustering algorithm itself [28,36], the number of clusters [12,49,57], or its initialization [26,37]. Instead of changing the clustering algorithm, a common method involves changing the data by randomly selecting sub-samples [11,32,38,48], using different features [23,37,43,51,52], employing random projections [8,37,39], or combining several methods together [47,53]. An alternative to the purely random approach generates the ensemble that maximizes a given criterion. First, a pool of partitions is created by using the strategies described above and a subset of this pool is then selected, which maximizes the objective function. For example, this greedy approach was used in [20], where a set of criteria were defined to obtain low, medium, and highly diverse ensembles.

These methods have been used widely to explore the dissimilarity within an ensemble, but the results obtained indicate that there is an important problem with current methods for ensemble diversity analysis. Indeed, previous studies provide opposing opinions regarding this issue, where some have suggested that more diverse ensembles are better for obtaining more accurate solutions [8,20], whereas others have proposed that moderate diversity is the preferred choice [15]. In addition to these contradictory results, high variability has been found not only among data sets but also when different ensemble generation strategies are employed. Moreover, plots of the accuracy as a function of diversity have shown that ensembles with similar diversity can differ greatly in their accuracy [15,20]. These confusing results show that current approaches can generate diversity but they cannot control it, and this limitation may lead to unpredictable outputs by the consensus method. This is an important issue and it must be addressed before any analysis of the impact of diversity on consensus clustering. This unpredictable behavior occurs because as one diversity measure is being observed, another properties of the ensemble are changing, thereby leading to erratic behavior by the consensus function. In addition, it is difficult to generate ensembles that are uniformly distributed in the diversity range under evaluation, which could lead to a biased analysis. Both of these reasons demonstrate the need to control the ensemble diversity in order to effectively analyze its impact on the consensus results.

Due to the importance of diversity in consensus clustering, the issues highlighted above motivated us to unveil a new problem in this area and to propose a novel method that allows fine-grained control of the ensemble diversity. To the best of our knowledge, no methods have been proposed previously for controlling disagreement among ensemble partitions. Our method extracts information from the ensemble structure and then uses it to make small changes, which decrease and increase the diversity among ensemble members in a smooth manner. The results that we obtained using six well-known data sets demonstrate that this method is effective for controlling the ensemble diversity, where the consensus function behaves in a smooth manner, thereby providing a novel approach for studying the impact of diversity on consensus clustering.

The remainder of this paper is organized as follows. In Section 2, we explain the problems with current methods and we define the steps in the diversity control method. Section 3 describes the data sets and performance measures used for testing. Section 4 presents the evaluation procedure and the results obtained. In Section 5, we summarize our conclusions as well as suggesting possible improvements and future research.

2. Novel method for controlling diversity

Current methods assume implicitly that ensembles with a particular level of diversity are comparable; thus, equal diversity values should represent similar ensembles, or at least similar inputs for the consensus function, which is expected to produce similar results. However, this might not be the case in practice. An example of such results is shown in Fig. 1, where the accuracy of the consensus partition [20] is plotted as a function of the pairwise ensemble diversity. Similar results can be found in [15]. Two problems are evident based on this plot. The first is the behavior of the output consensus accuracy (y-axis) when the pairwise diversity (x-axis) is around 0.21, where the diversity values are close to each other but many points differ greatly in their accuracy. A similar behavior can be observed around a diversity value of 0.28. Thus, ensembles with similar diversity can represent very different inputs for the consensus function. The second problem is that the diversity range is not always sampled uniformly; for example, there are far less ensembles with diversity values in [0.10, 0.20] and even none in [0.31, 0.38]. These two issues make the study of diversity a fairly difficult problem. As stated earlier,