Finding Alternative Clusterings

Jingyuan Zhang
OSCAR Team, SSDUT, DLUT
Sep. 19, 2010

Abstract

Traditional clustering techniques focus on producing only a single solution, while expecting the existence of alternative clusterings is quite reasonable since data in high dimensional space may have many alternative clusterings that exploit a different subset of features. Similarly, the underlying phenomena in the data could be quite complex and the data chosen to represent it is insufficient to justify one single explanation. So some approaches are proposed to find alternative clusterings.

One approach is to construct an algorithm with a **dual objective function** that simultaneously attempts to search for a different and good clusterings. It was taken by Bae and Bailey [1] in 2006 with considerable success. However, the dual objective function approach ties their approaches to a particular algorithm which may or may not be suitable for the task at hand. Another approach by Cui [2] in 2007 is to project the high dimensional data into an alternative **orthogonal subspace** but as we shall see this also has limitations such as not being appropriate for lower dimensional data sets such as spatial data. In order to make up for these two limitations, a third technique was introduced by Davidson and Qi [3] in 2008, which can be used with any number of clustering algorithms. It learns a distance function from the original clustering and performs a **liner transformation** to obtain a new data space. However, in many circumstances we may not wish to find a complete alternative, but perhaps a partial alternative, and seek to precisely state which parts of the clustering to retain and which parts not to retain. So a principled and flexible framework was created by Qi and Davidson [4] in 2009 as a solution to a **constrained optimization** problem that minimizes the difference between probability density functions of the original and a new transformed data set. The constraint on the optimization allows us to specify which properties of the clustering should be kept or not.

All these approaches were proposed to solve previous limitations in different ways. Maybe it is hard to improve the performance of these existing techniques or to find a new method much better. However, maybe we can consider whether these ideas can be applied to new data types such as uncertain data and stream data.

References

[1] E. Bae and J. Bailey. Coala: A novel approach for the extraction of an alternate clustering of high quality and high dissimilarity. In *ICDM '06*: *Proceedings of the Sixth International Conference on Data Mining*, pages 53–62, 2006.

[2] Y. Cui, X. Z. Fern, and J. G. Dy. Non-redundant multi-view clustering via orthogonalization. In *ICDM '07*: *Proceedings of the 2007 Seventh IEEE International Conference on Data Mining*, pages 133–142, 2007.

[3] I. Davidson and Z. Qi. Finding alternative clusterings using constraints. In *ICDM '08*: *Proceedings of the 2008 Eighth IEEE International Conference on Data Mining*, 2008.

[4] Z. Qi and I. Davidson. A Principled and Flexible Framework for Finding Alternative Clusterings. In *KDD '09*: *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2009.