MultiClust 2010: Discovering, Summarizing and Using Multiple Clusterings

Xiaoli Z. Fern
School of EECS
Oregon State Univeristy
Corvallis, OR
xfern@eecs.orst.edu

Ian Davidson
CS Department
University of California - Davis
Davis, CA
davidson@cs.ucdavis.edu

Jennifer G. Dy
ECE Department
Northeastern University
Boston, MA
jdy@ece.neu.edu

ABSTRACT

Traditional clustering focuses on finding a single best clustering solution from data. However, given a single data set, one could interpret it in different ways. This is particularly true with complex data that has become prevalent in the data mining community: text, video, images and biological data to name a few. It is thus of practical interest to find all possible alternative and interesting clustering solutions from data. Recently there has been increasing interest on developing algorithms to discover multiple clustering solutions from complex data. This report provides a description of the first international workshop on this emerging topic — SIGKDD MultiClust10: Discovering, Summarizing and Using Multiple Clusterings, which was held in Washington DC, on July 25th 2010. The workshop program consists of three invited talks and presentations of four full research papers and three short papers.

1. INTRODUCTION

Data is often multi-faceted by nature. Given a single data set, one can interpret it in several different ways. This is particularly true with complex data that has become prevalent in the data mining community: examples of such data include text, video, images and biological data. Yet, many data mining and clustering algorithms in particular only extract and present a single clustering/summarization even though multiple good alternatives exist. Practitioners oftentimes find that the clustering solution provided by an algorithm is not what they are looking for. Why limit the output to one clustering solution? Why not provide all possible alternative and interesting clustering solutions?

Recently, there has developed an emerging interest on discovering multiple clustering solutions from complex data. To avoid redundancy and excessive burden on the data analyst, it is key to extract clustering solutions that are informative yet non-redundant from one another. Toward this goal, important research issues include, how to define redundancy among clusterings, can existing algorithms be modified to accommodate this goal, how many solutions should we extract, how to select among exponentially many possible solutions which solutions to present to the data analyst, and how to most effectively help the data analyst find what he or she is searching for. Existing work approach this problem by looking for non-redundant, alternative, disparate or orthogonal clustering. Research in this area is developing and can benefit from well-established closely related areas, such as ensemble clustering and constraint-based clustering. This report presents a summary of the first Workshop on Discovering, Summarizing and Using Multiple Clusterings (MultiClust 10), which was held with ACM-SIGKDD 2010 in Washington, DC on June 25th, 2010. The workshop aims to bring together researchers from the above research areas to discuss important issues in multiple clustering discovery, compression and summarization. Our objectives are:

1. to further increase the general interest on this important topic in the broader research community;
2. to bring together experts from closely related areas (e.g., cluster ensembles and constraint-based clustering) to shed light on how this emerging new research direction can benefit from other well-established areas; and
3. to provide a venue for active researchers to exchange ideas and explore important research issues in this area.

In the remainder of the report, we present summary of the workshop program, which includes invited talks, and presentations of full research paper and short papers.

2. SUMMARY OF THE WORKSHOP

The half-day workshop was well attended and attracted 49 participants. The program included presentations of four full research papers and three short papers, as well as three talks from invited speakers: Prof. Joydeep Ghosh, of University of Texas at Austin, Prof. James Bailey of University of Melbourne, and Dr. Rich Caruana of Microsoft Research, and a panel discussion in the end. Below we provide a brief summary of the contributions. The detailed program, including the papers and presentation slides are accessible from the workshop website[1].

2.1 Invited Talks

The workshop had three invited speakers presenting different aspects of multiple clustering.

Prof. Joydeep Ghosh is an expert on clustering and has several published papers on ensemble clustering. In the workshop, he gave a thorough overview of ensemble methods for clustering and provided the audience with an insightful perspective of ensemble and multiple clustering.

[1] http://eecs.oregonstate.edu/research/MultiClust/

Prof. James Bailey presents a summary of the existing research on alternative clusterings. He organizes the existing techniques into different categories including sequential and simultaneous discovering of alternative clusterings and highlights the advantages/disadvantages of each type. He also notes two different styles in designing the algorithms—projection based methods and methods based on complex objective functions. Finally, Prof. Bailey also discusses important open issues in alternative clustering including challenges we face in evaluation and model selection.

Dr. Rich Caruana combines ensemble methods and the goal of generating multiple alternative clustering solutions in his work on meta clustering. In his talk, he compared two competing approaches for accomplishing the goal of efficiently finding multiple, significantly different, yet high quality clusterings, and to allow users to efficiently find among these the clustering(s) that are most useful for them. One approach is clustering with side information and the other is multi/meta clustering. One surprising result from their experiments is that the clustering which is most useful often is not a very compact clustering using common definitions of compactness.

2.2 Full Research Papers

We accepted four full research papers:

1. "Variational Inference for Nonparametric Multiple Clustering" by Y. Guan, J. Dy, D. Niu of Northeastern University and Z. Ghahramani of University of Cambridge;

2. "Uncovering Many Views of Biological Networks Using Ensembles of Near-Optimal Partitions" by G. Duggal, S. Navlakha, M. Girvan, and C. Kingsford of University of Maryland;

3. "Incorporating Spatial Similarity into Ensemble Clustering" by M. Ansari, N. Fillmore and M. Coen of University of Wisconsin-Madison; and

4. "On Using Class-Labels in Evaluation of Clusterings" by I. Färber, S. Günnemann, H.-P. Kriegel, P. Kröger, E. Müller, E. Schubert, T. Seidl, and A. Zimek of RWTH Aachen University and Ludwig-Maximilians-Univerrsität München, Germany.

Clustering is a difficult problem. This difficulty is compounded by that data may be multi-faceted. In addition, in high-dimensions, typically not all features are important. When designing feature selection algorithms for clustering, one needs to define a criterion for selecting which of two or more alternative feature subsets is the relevant/interesting subset. Why choose one feature subset, when all the alternative feature subset views might be interesting. Features irrelevant to one interpretation might be relevant to another interpretation. Guan, Dy, Niu and Ghahramani present a probabilistic nonparametric Bayesian model that can discover multiple clustering solutions and the feature subset views that generated each cluster partitioning simultaneously. They provide a variational inference approach to learn the features and clustering partitions in each view and also automatically learn the number of views and the number of clusters in each view.

Duggal, Navlakha, Girvan and Kingsford discuss finding multiple views of biological networks. This is one of the most developed application areas for alternative clustering. In particular they focus on densely interacting regions of biological networks. In this context producing one clustering reveals just a single view of how the cell is organized. The authors describe two approaches to show an ensemble of near-optimal partitions. They apply their work for a protein interaction network and show how their work can define robust communities. An important result found is that their solutions can genuinely represent alternative and complementary views of the networks' structure.

Ansari, Fillmore and Coen address a fundamental problem in ensemble clustering – namely, how should one compare the similarity of two clusterings? The vast majority of prior techniques for comparing clusterings are entirely partitional, i.e., they examine assignments of points in set theoretic terms after they have been partitioned. In doing so, these methods ignore the spatial layout of the data, disregarding the fact that this information is responsible for generating the clusterings to begin with. The authors demonstrate the importance of incorporating spatial information into forming ensemble clusterings and propose the use of the CDistance measure, which uses both spatial and partitional information to compare clusterings. They applied CDistance to address the correspondence problem, subsampling, stability analysis and diversity detection in ensemble methods.

The sound evaluation of clustering results in particular on real data is inherently difficult. In the literature, new clustering algorithms and their results are often externally evaluated with respect to an existing class labeling. These class-labels, however, may not be adequate for the structure of the data or the evaluated cluster model. Färber, Günnemann, Kriegel, Kröger, Müller, Schubert, Seidl and Zimek provided a survey of the literature on different related research areas that have observed this problem and discussed common "defects" that clustering algorithms exhibit with respect to this evaluation, and showed them on several real world data sets from different domains. They suggest that, a useful alternative evaluation method requires more extensive data labeling than the commonly used class labels or that it needs a combination of information measures to take subgroups, supergroups, and overlapping sets of traditional classes into account. They also initiated a discussion of the need for an evaluation scenario that regards the possible existence of several complementary sets of labels.

2.3 Short Research Papers

We accepted three short research and position papers:

1. "Less is More: Non-Redundant Subspace Clustering" by I. Assent of Aalborg University, Denmark, and E. Müller, S. Günnemann, R. Krieger and T. Seidl of RWTH Aachen University, Germany.

2. "Subspace Clustering, Ensemble Clustering, Alternative Clustering, Multiview Clustering: What Can We Learn From Each Other?" by P.-H. Kriegel and A. Zimek of Ludwig-Maximilians-Universität München, Germany.

3. "ASCLU: Alternative Subspace Clustering" by Stephan Günnemann, I. Färber, E. Müller and T. Seidl of RWTH Aachen University, Germany.

Assent, Müller, Gännemann, Krieger and Seidl present a position paper on identifying non-redundant, relevant subspace clusters. In particular the authors discuss techniques for evaluating and exploring subspace clusterings. They describe how their OpenSubSpace open source framework which contains implementations of various sub-space clustering algorithms also contains various measures of evaluating subspace clustering and their extensions for alternative sub-space clustering.

Kriegel and Zimek attempt to draw comparisons and similarities between the areas the workshop covers: subspace clustering, ensemble clustering, alternative clustering, and multiview clustering. They draw the conclusion that though superficially similar all are different approaches motivated by different problems and aiming at different goals. However, they do explore some connections between the areas and pose several topics for discussion amongst the fields:

- How do we treat diversity of clustering solutions? Under what conditions should diverse clusterings be unified or individually presented.
- How can we efficiently summarize diversity to an end-user
- How should we treat redundancy of clusters. Subspace clustering tries to get rid of too redundant clusters while alternative clustering allows some degree of redundancy. Is there a point where both research directions meet?
- How can we assess similarity between multiple clustering solutions? Various measures and indices have many undesirable traits.

Günnemann, Färber, Muller and Seidl present some of the earliest work on alternative sub-space clustering. This work differs from existing work, in that not only is the notion of alternativeness measured in terms of the cluster composition with respect to instances/objects but also to the sub-space the objects within the cluster occupy. An empirical analysis on several UCI data sets including the PenDigits data set shows the practicality of this work for even low dimensional data.

3. CONCLUSIONS

Discovering multiple clusterings from data is an emerging new topic that is gaining increasing interests among both researchers and practitioners. The MultiClust10 workshop is the first workshop on this topic. It brought together researchers from different subfields of data mining including cluster ensemble, constraint-based clustering and subspace clustering. The workshop provided an opportunity for researchers who are excited about this research area to communicate with one another and discuss different challenges in this emerging area. Significant interest has been expressed during the workshop to continue this effort and have another workshop for next year. More information about the MultiClust 10 workshop can be found on the following website, http://eecs.oregonstate.edu/research/MultiClust/.

4. ACKNOWLEDGMENTS

We would like to thank all the authors of the workshop papers, whose creative work made this workshop a success. We thank the invited speakers for sharing their insights on this important topic. We are greatly indebted to our excellent program committee members, who worked very hard to provide high quality reviews for the submissions. Specifically, the program committee members in alphabetical order are:

- Prof. Ira Assent of Aalborg University, Denmark
- Prof. James Bailey of University of Melbourne, Australia
- Prof. Arindam Banerjee of University of Minnesota
- Dr. Rich Caruana of Microsoft Research
- Prof. Chris Ding of University of Texas at Arlington
- Prof. Tao Li of Florida International University
- Emmanuel Muller of RWTH Aachen University, Germany
- Prof. Naren Ramakrishnan of Virginia Tech
- Prof. Thomas Seidl of RWTH Aachen University, Germany
- Dr. Alexander Topchy of Nielsen Media Research
- Dr. Kiri Wagstaff of the Jet Propulsion Lab of NASA.