Master Thesis Proposal

Learning with Total Graph Variation

Author:
Max Geiselbrechtinger, BSc.
Mat.Nr.:01609418

Supervisors:
Ao.Univ.Prof. Dipl.-Ing. Dr.techn. Gerald Matz
Univ.Ass. Dipl.-Ing. Thomas Dittrich BSc.

September 6, 2022

1 Motivation and Problem Statement

Graph structures naturally arise in applications of various domains such as biology, social sciences and logistics. The ability to construct graphs from pairwise relations between nodes makes them a flexible and powerful model for problems that exhibit complex structures. Another benefit of this local definition is its inherent scalability that enables the handling of large data quantities. For example complex social networks can be constructed solely from the information whether or not any two individuals are befriended. The resulting network can then be utilized to perform powerful inference tasks such as estimating political or consumer preferences or evaluating possible scenarios of epidemiological disease propagation. Another omnipresent task in machine learning is to separate networks into partitions that reflect similarity within and dissimilarity across them. This partitioning process is termed clustering or community detection and is substantial prerequisite in big data analysis [DM20].

Clustering on weighted graphs can be formulated as a minimum graph cut problem (ie. finding a set of edges with minimum weight that separates the cluster from the remaining graph). To avoid degenerate solutions (eg. clusters consisting of isolated vertices) usually an additional constraint that enforces reasonable cluster sizes is added to the problem. However, the resulting combinatorial minimization problem is NP-hard. In order to still be able to efficiently cluster large graphs a common approach is to relax the discrete problem setting to allow real valued cluster labels. The solution is then obtained by the eigenvectors of the relaxed problem and the methods are hence grouped under the term spectral clustering [VL07].

Despite its conceptually appealing formulation spectral clustering often yields results that differ significantly from the original combinatorial problem. This usually indicates a too loose relaxation and motivated the development of new algorithms which aim to minimize the total variation, a popular metric in image processing, of the label assignment. Total variation has the property of forming sharp label assignments in contrast to the squared differences metric in spectral clustering which promotes smooth transitions [BLUVB13].

Clustering as described above constitutes a strictly unsupervised scheme. However, it is also possible to incorporate prior knowledge in the form of initially labeled vertices (also called samples) to derive semi-supervised clustering algorithms [BNS04]. Such semi-supervised methods have become increasingly popular in recent years due to the rapid growth of unlabeled data collections accompanied by the lack of cheap labeling methods. Although numerous graph based semi-supervised clustering algorithms have been proposed, they are mostly restricted to similarity graphs (ie. graphs with non-negative edge weights). In some application dissimilarity provides valuable information that can not be exploited by unsigned graphical models. Classic examples that give rise to such opposing data relations are social networks or recommender systems. In the former users either follow or block each other whereas in the latter products can be liked or disliked by customers. In order to capture this kind of information semi-supervised learning algorithms have to be adapted to work on signed graphs [DM20].

2 Aim of the Work

In this thesis we focus on semi-supervised multi-class clustering of data that is represented by signed graphs. In particular we will aim to unify and build upon two different frameworks that leverage dissimilarity in data to find meaningful clusters within it. The first approach from Goldberg et. al [GZW07] originated from statistical learning theory and attempts to learn a classification function that fits the provided labels as well as obeys the dissimilarity structure of the signed graph. The second approach from Matz et. al [BDHM19] stems from the field of signal processing. Therein they define an extension of total variation to signed graphs and minimize the cluster label signal over this novel metric. By merging these two strategies from different field we attempt to combine the best aspects of both. In particular this work should reveal an efficient algorithm that is able to cope with generic data graphs, provides a tight fit to the underlying combinatorial problem, is able to generalize to unseen data and handle noisy class labels. The proposed scheme will be implemented with the Alternating Direction Method of Multipliers (ADMM) convex optimization algorithm that has seen increasing popularity in the machine learning community due to its ability to cope with large problems in a distributed fashion and to handle non-differential convex functions such as total variation [BPC$^+$11]. The implemented algorithm will be evaluated on synthetic data sets in order asses to its performance.

3 Methodological Approach

1. **Literature Review:**
   An extensive literature review will be conducted to acquire the necessary prerequisites in convex optimization as well as to examine the state-of-the-art of related approaches.

2. **Analytical derivation:**
   With tools from linear algebra and convex optimization the necessary update steps for the ADMM algorithm will be derived. This step should also reveal the connections between the two considered approaches.

3. **Algorithmic implementation:**
   The implementation of the ADMM algorithm will be performed in Python, which provides an extensive selection of machine learning and linear algebra packages that should aid the development.

4. **Numerical evaluation:**
   The implemented algorithms and their properties (parameter sensitivity etc.) will be evaluated on synthetic machine learning data sets.

4 State-of-the-Art

Clustering of data that exhibits network structure can be conveniently formulated as the minimization of graph cuts. However, practically meaningful formulations such as RatioCut or Ncut are in general NP-hard. The convex relaxation of such problems is termed spectral clustering, as it is based on the eigenvalues of graph related matrices (eg. the graph laplacian) [VL07].

Although most literature about spectral clustering is focused on similarity graphs there exist extensions to dissimilarity graphs as well. In particular Kunegis et. al [KSL$^+$10] derived a signed version of spectral clustering. In [DBM18] Dittrich et. al proposed a method for semi-supervised spectral clustering by incorporating previous knowledge of labels into the graphs weighted adjacency matrix.

Chiang et. al [CWD12] showed that signed spectral clustering cannot directly be generalized to case of multiple classes. Hence they derived the Balance Normalized Cut objective for k-way signed graph clustering.

In [CDGT19] they developed an algorithm that simultaneously minimizes and maximizes the normalized cut over the positive and negative graph edges respectively. To avoid the costly matrix inversion in the resulting objective function they resort to a generalized eigenproblem approach.

A semi-supervised machine learning method for multi-class clustering of unsigned graphs has been developed under the manifold regularization framework

by Beklin et. al [BNS04]. This framework utilizes unlabeled data for geometric regularization of popular supervised algorithms such as regularized least squares and support vector machines. An extension of this approach to account for dissimilarity graphs was proposed by Goldberg et. al in [GZW07].

The surge for a tighter relaxation of the underlying graph cut problem has led to a series of algorithms that utilize total variation, a popular measure in image processing. Bresson et. al in [BLUVB13] demonstrate the benefits of total variation clustering for unsigned graphs. Furthermore Matz et. al propose a signed version of the total variation and minimize it in a distributed fashion in [BDHM19].

5 Relevance to the Curriculum of Computer Engineering

This thesis will build upon and extend the theoretical as well as practical knowledge that has been taught in the curriculum of Computer Engineering. Besides the solid background and fundamentals that have been presented throughout this curriculum the modules Cyber-Physical Systems and Signal Processing will probably turn out to be most valuable for this work. They have introduced necessary prerequisites in linear algebra and statistical methods and sparked my interest in the field of machine learning. The following list contains courses from the curriculum which can be closely related to the topic of this thesis:

- 104.271 + 104.272 Discrete Mathematics
- 182.763 Stochastic Foundations of Cyber-Physical Systems
- 389.166 Signal Processing 1
- 389.170 Signal Processing 2
- 389.119 Parameter Estimation Methods
- 389.040 Signal Detection

References

[BDHM19] Peter Berger, Thomas Dittrich, Gabor Hannak, and Gerald Matz. Semi-supervised multiclass clustering based on signed total variation. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4953–4957. IEEE, 2019.

[BLUVB13] Xavier Bresson, Thomas Laurent, David Uminsky, and James Von Brecht. Multiclass total variation clustering. *Advances in Neural Information Processing Systems*, 26, 2013.

[BNS04] Mikhail Belkin, Partha Niyogi, and Vikas Sindhwani. Manifold regularization: A geometric framework for learning from examples. 2004.

[BPC+11] Stephen Boyd, Neal Parikh, Eric Chu, Borja Peleato, Jonathan Eckstein, et al. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine learning*, 3(1):1–122, 2011.

[CDGT19] Mihai Cucuringu, Peter Davies, Aldo Glielmo, and Hemant Tyagi. Sponge: A generalized eigenproblem for clustering signed networks. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1088–1098. PMLR, 2019.

[CWD12] Kai-Yang Chiang, Joyce Jiyoung Whang, and Inderjit S Dhillon. Scalable clustering of signed networks using balance normalized cut. In *Proceedings of the 21st ACM international conference on Information and knowledge management*, pages 615–624, 2012.

[DBM18] Thomas Dittrich, Peter Berger, and Gerald Matz. Semi-supervised spectral clustering using the signed laplacian. In *2018 52nd Asilomar Conference on Signals, Systems, and Computers*, pages 1413–1417. IEEE, 2018.

[DM20] Thomas Dittrich and Gerald Matz. Signal processing on signed graphs: Fundamentals and potentials. *IEEE Signal Processing Magazine*, 37(6):86–98, 2020.

[GZW07] Andrew B Goldberg, Xiaojin Zhu, and Stephen Wright. Dissimilarity in graph-based semi-supervised classification. In *Artificial Intelligence and Statistics*, pages 155–162. PMLR, 2007.

[KSL+10] Jérôme Kunegis, Stephan Schmidt, Andreas Lommatzsch, Jürgen Lerner, Ernesto W De Luca, and Sahin Albayrak. Spectral analysis of signed graphs for clustering, prediction and visualization. In *Proceedings of the 2010 SIAM international conference on data mining*, pages 559–570. SIAM, 2010.

[VL07] Ulrike Von Luxburg. A tutorial on spectral clustering. *Statistics and computing*, 17(4):395–416, 2007.