International Journal of Research
Available at https://edupediapublications.org/journals
p-ISSN: 2348-6848
e-ISSN: 2348-795X
Volume 03 Issue 12
August 2016

Discriminative Nonnegative Spectral Clustering With Flexible Constrained

[1]PEDDIREDDY KIRAN KUMAR REDDY, [2] Mrs.N.SUJATHA
[1] M.Tech (CSE) from JAGRUTI INSTITUTE OF ENGINEERING AND TECHNOLOGY
[2]Associate Professor, Department of Computer Science and Engineering, JAGRUTI INSTITUTE OF ENGINEERING AND TECHNOLOGY, Telangana State, India

ABSTRACT

We introduce an uncomplicated spectral approach to the well studied constrained clustering problem. This spectral approach captures constrained clustering as a generalized eigenvalue problem with graph Laplacians.And constrained clustered problem defined by three weighted graphs and these are the data graph, knowledge graph, disjoint graph. The algorithm works in nearly-linear time and provides concrete guarantees for the quality of the clusters, at least for the case of 2-way partitioning. In practice this translates to a very fast implementation that consistently outperforms existing spectral approaches both in speed and quality.

Keywords: Constrained spectral clustering, sparse coding, efficiency, scalability

1. INTRODUCTION

Clustering with constraints is a problem of central importance in machine learning and data mining. It captures the case when information about an application task comes in the form of both data and domain knowledge. We study the standard problem where domain knowledge is specified as a set of soft mustlink (ML) and cannot-link (CL) constraints [1]. The extensive literature reports a plethora of methods, including spectral algorithms that explore various modifications and extensions of the basic spectral. The distinctive feature of our algorithm is that it constitutes a natural generalization, rather than an extension of the basic spectral method. The generalization is based on a critical look at how existing methods handle constraints, in section 3. The solution is derived from a geometric embedding obtained via a spectral relaxation of an optimization problem, exactly in the spirit of . This is depicted in the workflow in Figure 1. Data and ML constraints are represented by a Laplacian matrix L and CL constraints by another Laplacian matrix H. The embedding is realized by computing a few eigenvectors of the generalized eigenvalue problem $Lx = \lambda Hx$. The generalization of lies essentially in H being a Laplacian matrix rather than the diagonal D of L. In fact, as we will discuss later, D itself is equivalent to a specific Laplacian matrix; thus our method encompasses the basic spectral method as a special case of constrained clustering. Figure 1: A schematic overview of our approach. Our approach is characterized by its conceptual simplicity that enables a straightforward mathematical derivation of the algorithm, possibly the simplest among all competing spectral methods. Reducing the problem to a relatively simple generalized eigensystem enables us to derive directly from

recent significant progress due to Lee et al. in the theoretical understanding of the standard spectral clustering method, offering its first practical realization. In addition, the algorithm comes with two features that are not simultaneously shared by any of the prior methods: (i) it is provably fast by design as it leverages fast linear system solvers for Laplacian systems [9] (ii) it provides a concrete theoretical guarantee for the quality of 2-way constrained partitioning, with respect to the underlying discrete optimization problem, via a generalized Cheeger inequality (section 6). In practice, our method is at least 10x faster than competing methods on large data sets. It solves data sets with millions of points in less than 2 minutes, on very modest hardware. Furthermore the quality of the computed segmentations is often dramatically better.

2. Related Work

The use of supervision in clustering tasks has been addressed in many ways. A first related approach is that of [33], which is inspired by distance learning. Constraints are given through a set of data point pairs that should be close. The authors then consider the problem of learning a weighted matrix of similarities. They derive an optimization problem of high complexity, which they solve by doing alternate gradient ascent on two objectives, one bringing closer points that are similar and the other putting off the other points. Similarly, in learning spectral clustering is the problem of finding weighted matrix or the spectrum of the Gram matrix given a known partition. A related field is supervised clustering [9], the problem of training a clustering algorithm to produce desirable clusterings: given sets of items and complete clusterings over these sets, we learn how to cluster future sets of items.

Another set of related approaches are constrained versions of the k-means clustering algorithm. it is proposed that, at each step of the algorithm, each point is assigned to the closest centroid provided that must-link and cannotlink constraints are not violated. It is not clear how the choice of the ordering on points affects the clustering. Moreover, constraints are considered as hard constraints which makes the approach prone to noise effects. Kulis et al improve on the work of in . Their algorithm relies on weighted kernel k-means ([8]). The authors build a kernel matrix $K = \sigma I + W + S$, where W is a similarity matrix, S is a supervision matrix such that Sij is positive (respectively negative) when nodes i and j must link (respectively cannot link) or zero when unconstrained. The addition of σI ensures the positive semi-definiteness of K (otherwise, K would not be a kernel, would not have any latent Euclidean space, a requirement for k-means to converge and for theoretical justification). Introducing constraints in spectral clustering has received a lot of attention in the last decade ([34,14,5,19,32]). In many cases, the proposed approaches rely on a modification of the similarity matrix and then the resolution of the associated approximated normalized cut. For instance, in , weights in the similarity matrix are forced to 0 or 1 following must-link and cannot-link constraints. But this kind of weights may have a limited impact on the result of the clustering, in particular when the considered two nodes have many paths that link them together. consider three kinds of constraint and cast them into an optimization problem including membership constraints in a 2-partitioning graph problem. To guarantee a smooth solution, they reformulate the optimization problem so that it involves computing the eigen decomposition of the graph Laplacian associated with the data.

The approach relies on an optimization procedure that includes nullity of the flow from labeled nodes in cluster 1, to labeled nodes in cluster 2. The algorithm closely resembles the semi-supervised harmonic Laplacian approach developed for instance in . But this approach is also limited to the binary case. In , pairwise constraints are used to propagate affinity information to the other edges in the graph. A closed form of the optimal similarity matrix can be computed but its computation requires one matrix inversion per cannot-link constraint. In , constrained clustering is done by learning a transformation of the spectral embedding into another space defined by a kernel. The algorithm attempts to project data points representing nodes onto the bound of a unithypersphere. The inner product between vectors describing nodes that must link is close to 0, and the inner product between vectors describing nodes that cannot-link is close to 1. That way, if a node vi belongs to the cluster j, then the vector vi describing vi will be projected to ∗j where ej is a vector of length k full of zeros except on the jth component where it is equal to 1. The number of dimensions of the hypersphere is directly related to the ability to separate clusters. One drawback is that the algorithm uses semidefinite programs whose size is quadratic in that number of dimensions. Recently, propose to include constraints by modifying directly the optimization problem rather than modifying the Laplacian. In their algorithm called csp, they introduce a matrix Q where Qij is 1 if i and j must-link, −1 if i and j cannot-link and 0 otherwise. Then, a constraint f >Qf > α is added to the normalized cut objective considered in unconstrained spectral clustering. Parameter α is considered as a way to soften constraints. Their approach outperforms previous approaches such as the one based on kernel k-means defined in .An original approach based on tight relaxation of graph cut is presented in . The approach deals with must and cannot-links but in the two clusters case. It guarantees that no constraints are violated as long as they are consistent. For problems with more than two clusters, hierarchical clustering is proposed. Unfortunately in this case, the algorithm loses most of its theoretical guarantees.

3. SPECTRAL CLUSTERING ALGORITHM

Spectral clustering is is appealingly simple: Given some data, you build an affinity (or kernel) matrix, analyze its spectrum, and often get a perfect clustering from the dominant eigen vectors for free. This simple algorithm[4] or its slightly more complex variants which yield so good results are widely appreciated for applications. Here are the key steps of spectral clustering algorithm: Given a set of points S = {s1, . . . , sn } in a high dimensional space. 1. Form a distance matrix D ∈ R2 . This distance measure is Euclidean, but other measures also make sense. 2. Transform the distance matrix to an affinity matrix by Aij = exp(− σij) if i ≠ j,0 if i = j. The free parameter σ 2 controls the rate at which affinity drops off with distance. 3. Form the diagonal matrix D whose (i,i) element is the sum of A's ith row, and create the Laplacian matrix L = D−1/2 AD−1/2 4. Obtain the eigenvectors and eigenvalues of L. 5. Form a new matrix from the vectors associated with the k largest eigenvalues. Choose k by using eigen gap method. 6. Each item now has a vector of k coordinates in the transformed space. Normalize these vectors to unit length. 7. Cluster in k-dimensional space. The result will be k well

International Journal of Research
Available at https://edupediapublications.org/journals
p-ISSN: 2348-6848
e-ISSN: 2348-795X
Volume 03 Issue 12
August 2016

separated clusters. Spectral clustering is a more advanced algorithm compared to k-means as it uses several mathematical concepts (i.e. degree matrices weight matrices, similarity matrices, similarity graphs, graph Laplacians, eigenvalues and eigenvectors) in order to divide similar data points in the same group and dissimilar data points in different groups. This Spectral Clustering works well for many real world data sets eventhough, it needs some modification in terms of improving its time complexity, space complexity.

4. Re-thinking constraints

Many approaches have been pursued within the constrained spectral clustering framework. They are quite distinct but do share a common point of view: constraints are viewed as entities structurally extraneous to the basic spectral formulation, necessitating its modification or extension with additional mathematical features. However, a key fact is overlooked: Standard clustering is a special case of constrained clustering with implicit soft ML and CL constraints. To see why, let us briefly recall the optimization problem in the standard method (Ncut). $\varphi = \min_{S \subseteq V} \frac{\text{cutGD}(S, S^-)}{\text{vol}(S)\text{vol}(S^-)/\text{vol}(V)}$

Here vol(S) denotes the total weight incident to the vertex set S, and cutG(S, S^-) denotes the total weight crossing from S to S^- in G. The data graph GD is actually an implicit encoding of soft ML constraints. Indeed, pairwise affinities between nodes can be viewed as 'soft declarations' that such nodes should be connected rather than disconnected in a clustering. Let now di denote the total incident weight of vertex i in GD. Consider the demand graph K of implicit soft CL constraints, de- fined by the adjacency

Kij = didj/vol(V). It is easy to verify that vol(S)vol(S^-)/vol(V) = cutK(S, S^-). We have min S⊆V cutGD (S, S^-) vol(S)vol(S^-)/vol(V) = min S⊆V cutGD (S, S^-) cutK(S, S^-) . In other words, the Ncut objective can be viewed as: min S⊆V weight of cut (violated) implicit ML constraints weight of cut (satisfied) implicit CL constraints . (1) With this realization, it becomes evident that incorporating the knowledge graphs (GML, GCL) is mainly a degree-of-belief issue, between implicit and explicit constraints. Yet all existing methods insist on handling the explicit constraints separately. For example, modify the Ncut optimization function by adding in the numerator the number of violated explicit constraints (independently of them being ML or CL), times a parameter γ. In another example, solve the spectral relaxation of Ncut, but under the constraint that the number of satisfied ML constraints minus the number of violated CL constraints is lower bounded by a parameter α. Despite the separate handling of the explicit constraints, degree-ofbelief decisions (reflected by parameters α and γ) are not avoided. The actual handling also appears to be somewhat arbitrary. For instance, most methods take the constraints unweighted, as usually provided by a user, and handle them uniformly; but it is unclear why one constraint in a densely connected part of the graph should be treated equally to another constraint in a less well-connected part. Moreover, most prior methods enforce the use of the balance implicit constraints in K, without questioning their role, which may be actually adverserial in some cases. In general, the mechanisms for including the explicit constraints are oblivious of the input, or even of the underlying algebra. Our approach.

We choose to temporarily drop the distinction of the constraints into explicit and implicit. We instead assume that we are given one set of ML constraints, and one set of CL constraints, in the form of weighted graphs G and H. We then design a generalized spectral clustering method that retains the k-way version of the objective shown in equation 1. We apply this generalized method to our original problem, after a merging step of the explicit and implicit CL/ML constraints into one set of CL/ML constraints. The merging step can be left entirely up to the user, who may be able to exploit problem-specific information and provide their choice of weights for G and H. Of course, we expect that in most cases explicit CL and ML constraints will be provided in the form of simple unweighted graphs GML and GCL. For this case we provide a simple method that resolves the degree-of-belief issue and constructs G and H automatically. The method is heuristic, but not oblivious to the data graph, as they adjust to it.

5. CONCLUSION

There are several methods implemented in our approach Our method proposes a natural way (based on a Gaussian process formulation) to propagate affinity information through pair wise constraints; the latter act as wormholes that connect space regions that are faraway (low affinity) for must-links, or disconnect nearby regions for cannot links. The new affinity matrix has a closed-form expression (eq. 2–3) that can be obtained by inverting a small matrix, at a negligible overhead over spectral clustering. This new affinity can be represented by a new kernel function derived from the original one. Experimentally, our method needs very few constraints to achieve good clustering's as compared with other methods.

REFERENCES:

[1] S. Basu, M. Bilenko, and R. Mooney. A probabilistic framework for semi-supervised clustering. In KDD, 2004.

[2] H. Chang and D. Yeung. Semi-supervised metric learning by kernel matrix adaptation. In Int. Conf. Machine Learning and Cybernetics (ICMLC'05), 2005.

[3] C. Chennubhotla and A. Jepson. EigenCuts: Half lives of eigenflows for spectral clustering. In NIPS, 2003.

[4] W. Chu, V. Sindhwani, Z. Ghahramani, and S. Keerthi. Relational learning with Gaussian processes. In NIPS, 2007.

[5] J. Davis, B. Kulis, P. Jain, S. Sra, and I. Dhillon. Informationtheoretic metric learning. In ICML, 2006.

[6] Jaya Kawale and Daniel Boley. Constrained spectral clustering using l1 regularization. In SDM, pages 103–111, 2013.

[7] Andrew V. Knyazev. Toward the optimal preconditioned eigensolver: Locally optimal block preconditioned conjugate gradient method. SIAM J. Sci. Comput, 23:517–541, 2001.

[8] Ioannis Koutis, Gary L. Miller, and Richard Peng. A nearly m log n solver for SDD linear systems. In FOCS '11: Proceedings of the 52nd Annual IEEE Symposium on Foundations of Computer Science, 2011.

[9] Ioannis Koutis, Gary L. Miller, and Richard Peng. A fast solver for a class of linear systems. Commun. ACM, 55(10):99–107, October 2012.

[10] Ohad Shamir, Naftali Tishby; Spectral Clustering on a Budget. JMLR W&CP 15:661-669, 2011. Retrieved from http://www.jmlr.org/proceedings/papers/v15/shamir11a/shamir11a.p df

[11] Fabian L. Wauthier, Nebojsa Jojic and Michael I. Jordan, Active Spectral Clustering via Iterative Uncertainty Reduction. KDD '12 Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining Pages 1339-1347. Retrieved from http://www.stats.ox.ac.uk/~wauthier/wauthier-jojicjordan-kdd12.pdf

Peddireddy kiran kumar reddy: is pursuing M.Tech degree in, Computer Science and Engineering from Jagruti Institute of Engineering and Technology, Telangana State, India.

Mrs.N.SUJATHA is presently working as Associate Professor in, Department of computer science and engineering, Telangana State,India.She has published several research papers in both International and National conferences and Journals.