SPECTRAL CLUSTERING WITH IMBALANCED DATA

Jing Qian *Venkatesh Saligrama*

Boston University
Boston, MA, 02215

ABSTRACT

Spectral clustering is sensitive to how graphs are constructed from data. In particular, if the data has proximal and imbalanced clusters, spectral clustering can lead to poor performance on well-known graphs such as k-NN, ϵ-neighborhood and full-RBF graphs. We propose a graph partitioning problem that seeks minimum cut partitions under minimum size constraints on clusters to deal with imbalanced data. Our approach parameterizes a family of graphs by adaptively modulating node degrees on a fixed node set, to yield a set of parameter dependent cuts reflecting varying levels of imbalance. The solution to our problem is then obtained by optimizing over these parameters. We present asymptotic limit cut analysis to justify our approach. Experiments on synthetic and real data sets demonstrate the superiority of our method.

Index Terms— Spectral Clustering; Imbalanced Data; RatioCut/Normalized Cut

1. INTRODUCTION

Data with imbalanced clusters arises in many learning applications and has attracted much interest [1]. In this paper we focus on graph-based spectral methods for clustering tasks. In spectral methods, a graph representing data is first constructed and then spectral clustering (SC) [2, 3] is applied on the graph. Common graph construction methods include ϵ-graph, fully-connected RBF-weighted(full-RBF) graph and k-nearest neighbor (k-NN) graph. Of the three k-NN graphs appears to be most popular due to its relative robustness to outliers [4, 5]. We show that the poor performance of spectral methods on imbalanced data can be attributed to applying Ratio-Cut (RCut) or normalized cut (NCut) minimization objectives on traditional graphs, which sometimes tend to emphasize balanced partition size over small cut-values.

To deal with imbalanced data we propose partition constrained minimum cut problem (PCut). Size-constrained min-cut problems appear to be computationally intractable [6, 7]. Instead we attempt to solve PCut on a parameterized family of cuts. To realize these cuts we parameterize a family of graphs over some parametric space $\lambda \in \Lambda$ and generate candidate cuts using spectral methods as a black-box. This requires a sufficiently rich graph parameterization capable of approximating varying levels of imbalanced data. To this end we introduce a novel parameterization for graphs that adaptively modulates node degrees in varying proportions. We then solve PCut on a baseline graph over the candidate cuts obtained from this parameterization. Fig. 1 depicts our approach for binary clustering. Our limit cut analysis shows that our approach asymptotically does adapt to imbalanced clusters. We demonstrate the superiority of our method through unsupervised clustering experiments on synthetic and real data sets.

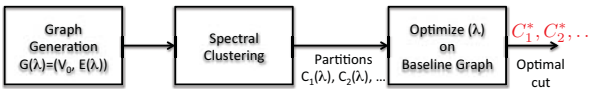


Fig. 1. Proposed Framework for Clustering on Imbalanced Data.

Related Work:
Sensitivity of spectral methods to graph construction is well documented [5, 8, 9]. [10] suggests an adaptive RBF parameter in full-RBF graphs to deal with imbalanced clusters. [11] describes these drawbacks from a random walk perspective. [12, 13] also mention imbalanced clusters, but none of these works explicitly deal with imbalanced data. Besides, our approach is complementary to their schemes and can be used in conjunction. Another related approach is size-constrained clustering [14, 15, 16, 17, 18, 6], which is shown to be NP-hard. [19] proposes sub-modularity based schemes that work only for some special cases. Besides, these works either impose exact cardinality constraints or upper bounds on the cluster sizes to look for balanced partitions. While this is related, we seek minimum cuts with lower bounds on smallest-sized clusters. Minimum cuts with lower bounds on cluster size naturally arises because we seek cuts at density valleys (accounted for by the min-cut objective) while rejecting singleton clusters and outliers (accounted for by cluster size constraint).

The organization of the paper is as follows. In Section 2 we propose our partition constrainted min-cut (PCut) framework and describe our algorithm. We explore the theoretical analysis in Section 3. In Section 4 we present clustering experiments on synthetic and real data sets to show significant improvements when data is imbalanced.

978-1-4799-2893-4/14/$31.00 ©2014 IEEE

2. PARTITION CONSTRAINED MIN-CUT (PCUT)

Assume that data is drawn IID from some unknown density $f(x)$, where $x \in \mathbf{R}^d$. Let $G = (V, E, W)$ be a weighted undirected graph constructed from n samples. Each node $v \in V$ is associated with a data sample. Edges are constructed using one of several graph construction techniques such as k-NN graph. The weights on the edges are similarity measures such as RBF kernels that are based on Euclidean distances. We denote by S a cut that partitions V into C_S and $\bar{C}_S$. The cut-value associated with S is:

$$Cut(C_S, \bar{C}_S) = \sum_{u \in C_S, v \in \bar{C}_S, (u,v) \in E} w(u, v) \tag{1}$$

For binary clustering tasks, the aim is to seek a hypersurface S that samples within each cluster resulted from S are similar to each other while dissimilar across clusters. The simple min-cut framework directly aims at low-density cuts, but is well known to be vulnerable to outliers. Spectral clustering attempts to minimize RCut/NCut objectives:

$$\min_S : \ Cut(C_S, \bar{C}_S) \left(\frac{size(V)}{size(C_S)} + \frac{size(V)}{size(\bar{C}_S)} \right), \tag{2}$$

where $size(C) = \sum_{u \in C, v \in V} w(u, v)$ for NCut and $size(C) = |C|$ for RCut. Both objectives seek to trade-off small cut-values against cut size. While robust to outliers, minimizing RCut (NCut) can lead to poor performance when data is imbalanced, as will be shown in experiments.

We now propose our partition constrained min-cut (PCut) problem in finite data setting, which seeks low-density cuts with lower bounds on the size of the smallest cluster obtained:

$$S_* = \arg\min_S \left\{ Cut(C_S, \bar{C}_S) \mid \min\{|C_S|, |\bar{C}_S|\} \geq \delta |V| \right\}. \tag{3}$$

Eq.(3) describes a binary partitioning problem but generalizes to arbitrary number of partitions. Note that different from other size-constrained partitioning methods [16, 17, 18] which either focus on balanced partitions or cuts with exact cardinality constraints, we aim to identify natural low-density cuts that are not too small. Various versions of this problem is known to be NP-hard [7]. We here employ SC as a black-box to generate candidate cuts on a suitably parameterized family of graphs. Specifically, our idea is a parameterization that selectively removes/adds edges in low/high density regions, inducing a much smaller cut value at density valleys. This is achieved by modulating node degrees based on ranking of data samples, which reflects the relative density at nodes. We call the resulting graph the Rank-Modulated Degree (RMD) graph. Eq.(3) is then optimized over these candidate cuts.

2.1. PCut: Algorithms

Given n data samples, our task is unsupervised clustering, assuming the number of clusters/classes K is known. We start with a baseline k_0-NN graph $G_0 = (V, E_0)$ built on samples with k_0 large enough to ensure graph connectivity. Main steps of our PCut are as follows.

Main Algorithm: RMD Graph-based PCut

1. Compute the rank $R(x_i)$ of $x_i, i = 1, ..., n$;
2. For different parameter configurations,
   a. Construct the parametric RMD graph;
   b. Apply spectral clustering to obtain a K-partition on the current RMD graph;
3. Among various partition results, pick the "best".

(1) Rank Computation:
We compute the rank $R(v)$ of every node v as follows:

$$R(x_v) = \frac{1}{n} \sum_{w \in V} \mathbf{I}_{\{\eta(x_v) \leq \eta(x_w)\}} \tag{4}$$

where $\mathbf{I}$ denotes the indicator function, $\eta(x_v)$ is some statistic reflecting the relative density at v. We here choose average nearest neighbor distance, where $N(v)$ is the set of neighbors of v on the baseline graph:

$$\eta(x_v) = \frac{1}{|N(v)|} \sum_{w \in N(v)} \|x_v - x_w\|. \tag{5}$$

It is shown that this statistic outperforms other choices such as single k-NN distance or ϵ-neighborhood density in [20].
(2) Parameterized family of RMD graphs:
We connect each node v with its $k_\lambda(v)$ nearest neighbors:

$$k_\lambda(v) = k(\lambda + 2(1 - \lambda)R(x_v)), \tag{6}$$

This generates RMD parameterization. In experiments we also vary k and RBF parameter σ for a richer family of RBF-RMD graphs. Let $G(\lambda, k, \sigma)$ be the generated graph with parameter λ, k, σ.
(3) Parameterized family of cuts:
Spectral clustering is applied on $G(\lambda, k, \sigma)$. We thus obtain a family of K-partitions: $C_1(\lambda, k, \sigma)$, $C_2(\lambda, k, \sigma)$, ..., $C_K(\lambda, k, \sigma)$.
(4) Parameter Optimization:
The final step is to solve Eq.(3) on the baseline graph G_0. We assume prior knowledge that we want partitions of which the smallest cluster is at least of size δn.

$$\begin{aligned} &\min_{\lambda,k,\sigma} \{Cut_0\,(C_1, ..., C_K) = \textstyle\sum_{i=1}^K Cut_0(C_i, \bar{C}_i)\} \\ &s.t. \ \ \min\{|C_1(\lambda, k, \sigma)|, ..., |C_K(\lambda, k, \sigma)|\} \geq \delta n \end{aligned} \tag{7}$$

$Cut_0(\cdot)$ denotes evaluating cut values on the baseline graph G_0. Partitions with clusters smaller than δn are discarded.
Remark:
1. Although step (4) suggests a grid search over several parameters, it turns out that other parameters such as k, σ do not play an important role as λ. Indeed, Sec.4 will show that

while step (4) can select appropriate k, σ, it is by searching over λ that adapts spectral clustering to data with varying levels of imbalancedness (also see Thm.2).
2. Our framework uses existing spectral algorithms and so can be combined with other graph-based partitioning algorithms to improve performance for imbalanced data, such as 1-spectral clustering, sparsest cut or minimizing conductance [12, 21, 22, 23].

3. ANALYSIS

To justify our method, we establish asymptotic consistency of ranks and the limit cut behavior of spectral clustering on RMD graph. Due to space limit we omit the proofs.

Suppose $f(\cdot)$ has a compact support and is continuous and bounded: $f_{max} \geq f(x) \geq f_{min} > 0$. It is smooth, i.e. $||\nabla f(x)|| \leq \lambda$, where $\nabla f(x)$ is the gradient of $f(\cdot)$ at x. There is no flat regions, i.e. $\forall \sigma > 0$, $\mathcal{P}\{y : |f(y) - f(x)| < \sigma\} \leq M\sigma$ for all x in the support, where M is a constant.

First we show the rank $R(y)$ at some point y converges to the p-value function $p(y)$. Note that p exactly follows the shape of f and always ranges in $[0, 1]$ no matter how f scales.

Theorem 1. *Assume $f(x)$ satisfies the above regularity conditions. As $n \to \infty$, we have*

$$R(y) \to p(y) := \int_{\{x: f(x) \leq f(y)\}} f(x)dx. \tag{8}$$

Next we study the limit behavior of RCut (NCut) induced on unweighted RMD graph. Assume for simplicity that each node v is connected to exactly $k_\lambda(v)$ nearest neighbors of Eq.(6).

Theorem 2. *Assume f satisfies the above regularity conditions and also the general assumptions in [8]. S is a fixed hyperplane in $\boldsymbol{R}^d$. Unweighted RMD graph is generated according to Eq.(6), where $\lambda \in (0, 1)$ is a constant. Let $\rho(x) = \lambda + 2(1-\lambda)p(x)$. Assume $d \geq 2$, $k_n/n \to 0$ and $k_n/\log n \to \infty$. Then as $n \to \infty$ we have that:*

$$\frac{1}{k_n}\sqrt[d]{\frac{n}{k_n}}RCut_n(S) \longrightarrow C_d B_S \int_S f^{1-\frac{1}{d}}(s)\rho^{1+\frac{1}{d}}(s)ds. \tag{9}$$

$$\sqrt[d]{\frac{n}{k_n}}NCut_n(S) \longrightarrow C_d B_S \int_S f^{1-\frac{1}{d}}(s)\rho^{1+\frac{1}{d}}(s)ds. \tag{10}$$

where $C_d = \frac{2\eta_{d-1}}{(d+1)\eta_d^{1+1/d}}$, $B_S = \left(\mu(C^+)^{-1} + \mu(C^-)^{-1}\right)$, and $\mu(C^\pm) = \int_{C^\pm} f(x)dx$.

Remark: In the limit cut behavior, without ρ term, the balancing term $B_S = 1/\alpha(1-\alpha)$ could induce a larger RCut (NCut) value for density valley cut than balanced cut when the underlying data is imbalanced. Our RMD scheme appends $\rho(s) = (\lambda + 2(1-\lambda)p(s))$, which is monotonic in p-value. So the cut-value at low/high density regions can be significantly reduced/increased. Indeed for small λ value, cuts S near peak densities have $p(s) \approx 1$ and so $\rho(s) \approx (2)^{1+\frac{1}{d}}$, while near valleys we have $\rho(s) \approx (\lambda)^{1+\frac{1}{d}} \ll 1$.

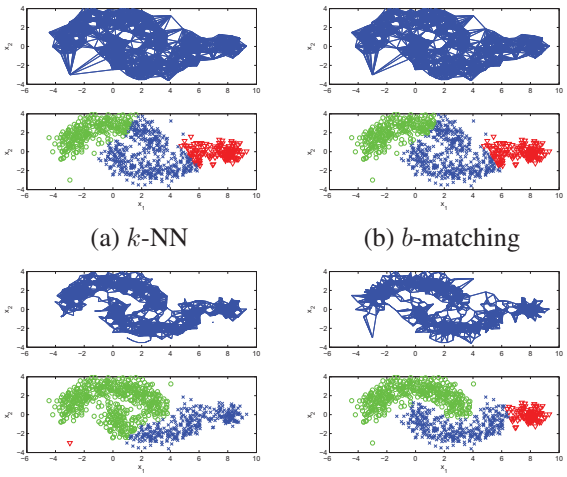
4. EXPERIMENTS

We carry out unsupervised clustering experiments on both synthetic and real data sets. We focus on imbalanced data by randomly sampling from different classes disproportionately. We compare our RMD graph with full-RBF, ϵ-graph, RBF k-NN, b-matching graph [9] and full graph with adaptive RBF (full-aRBF) [10], all conducted within our PCut framework. We view each as a family of graphs parameterized by their relevant parameters and optimize over different parameters as described in Sec. 2.1 and Eq.(7). Error rates are averaged over 20 trials.

Time Complexity: RMD graph construction is $O(dn^2 logn)$ (similar to k-NN graph). Computing cut value and checking cluster size for a partition takes $O(n^2)$. So if totally D graphs are parameterized; complexity of learning algorithm is T, the time complexity is $O(D(dn^2 logn + T))$.

Tuning Parameters: Note that the only parameters left are: (a) k_0 in the baseline graph. This is fixed to be $\sqrt{n}$. (b) Size threshold δ. We fix this a priori to be about 0.05, i.e., 5% of all samples.

Evaluation against Oracle: To evaluate the effectiveness of our framework (Fig.1) and RMD parameterization, we compare against an ORACLE that is tuned to both ground truth labels as well as imbalanced proportions.



(a) k-NN (b) b-matching (c) ϵ-graph(full-RBF) (d) RMD

Fig. 2. Results of 3-way RCut-based spectral clustering on 2 moons and 1 gaussian component data set.

Synthetic Illustrative Example
Consider a multi-cluster complex-shaped data set consisting of 1 small Gaussian and 2 moon-shaped proximal clusters in

Table 1. Error rates of normalized SC on various graphs for imbalanced real data sets. First row ("BO" Balanced Oracle) shows RBF k-NN results on imbalanced data with k, σ tuned using ground truth labels but on balanced data. Last row ("O" Oracle) shows the best ORACLE results of RBF RMD on imbalanced data.

Error Rates (%)	USPS		SatImg			OptDigit			LetterRec	
	8vs9	1,8,3,9	4vs3	3,4,5	1,4,7	9vs8	6vs8	1,4,8,9	6vs7	6,7,8
RBF k-NN (BO)	33.20	17.60	15.76	22.08	25.28	15.17	11.15	30.02	7.85	38.70
RBF k-NN	16.67	13.21	12.80	18.94	25.33	9.67	10.76	26.76	4.89	37.72
RBF b-match	17.33	12.75	12.73	18.86	25.67	10.11	11.44	28.53	5.13	38.33
full-RBF	19.87	16.56	18.59	21.33	34.69	11.61	15.47	36.22	7.45	35.98
full-aRBF	18.35	16.26	16.79	20.15	35.91	10.88	13.27	33.86	7.58	35.27
RBF RMD	4.80	9.66	9.25	16.26	20.52	6.35	6.93	23.35	3.60	28.68
RBF RMD (O)	3.13	7.89	8.30	14.19	18.72	5.43	6.27	19.71	3.02	25.33

Fig.2. Sample size $n = 1000$ with the rightmost small cluster 10% and two moons 45% each. This example is only for illustrative purpose with a single run, so we did not parameterize the graph or apply step (4). We fix $\lambda = 0.5$, and choose $k = l = 30$, $\epsilon = \sigma = \tilde{d}_k$, where $\tilde{d}_k$ is the average k-NN distance. Model-based approaches can fail on such dataset due to the complex shapes of clusters. The 3-way SC based on RCut is applied. On k-NN and b-matching graphs SC fails for two reasons: (1) SC cuts at balanced positions and cannot detect the small cluster; (2) SC cannot recognize the winding low-density regions between 2 moons due to too many spurious edges. SC fails on ϵ-graph (similar on full-RBF) because the outlier point forms a singleton cluster, and also cannot recognize the low-density curve. While robust to outliers, our RMD graph significantly sparsifies the graph at low-density regions, enabling SC to cut along the valley and detect the small cluster.

Real Experiments

We focus on imbalanced settings and apply normalized spectral clustering on several real datasets including USPS (256-dim), Statlog landsat satellite images (4-dim), letter recognition images (16-dim) and optical recognition of handwritten digits (16-dim) [24]. We construct k-NN, b-match, full-RBF and RMD graphs all combined with RBF weights, but do not include ϵ-graph due to its overall poor performance [9].

Our sample size varies from 750 to 1500. We discretize not only λ but also k, σ to parameterize graphs. We vary k in $\{5, 10, 20, 30, \ldots, 100, 120, 150\}$. While small k may lead to disconnected graphs this is not an issue since singleton cluster candidates are ruled infeasible in our PCut framework. Also notice that for $\lambda = 1$, RMD graph is identical to k-NN graph. For RBF parameter σ it has been suggested to be of the same scale as the average k-NN distance $\tilde{d}_k$ [25]. This suggests a discretization of σ as $2^j \tilde{d}_k$ with $j = -3, -2, \ldots, 3$. We discretize $\lambda \in [0, 1]$ and varied in steps of 0.2. Data sets are sampled in an imbalanced way shown in Tab.2.

In Tab.1 the first row is the imbalanced results of RBF k-NN using ORACLE k, σ parameters tuned with ground-truth labels on balanced data for each data set (300/300, 250/250/250, 250/250/250/250 samples for 2,3,4-class cases). Comparison of first two rows reveals that the ORACLE choice on balanced data may not be suitable for imbalanced data, while our PCut framework, although agnostic of the labels, picks more suitable parameters k, σ for RBF k-NN graph.

Table 2. Imbalancedness of data sets.

Data sets	#samples per class
2-class (eg. USPS 8vs9)	150/600
3-class (eg. SagImg 3/4/5)	200/400/600
4-class (eg. USPS 1/8/3/9)	200/300/400/500

The last row presents ORACLE results on our RBF-RMD graph tuned to imbalanced data. Comparison of the last two rows shows that our PCut on RMD, agnostic of true labels, closely approximates the oracle performance. Furthermore, Tab.1 demonstrates that our RMD graph parameterization equipped with PCut framework performs consistently better than other methods.

5. CONCLUSION

In this paper we propose the partition constraint min-cut (PCut) framework, which seeks min-cut partitions under minimum cluster size constraints. Since constrained min-cut is NP-hard, we adopt existing spectral methods (SC, GRF, GTAM) as a black-box subroutine on a parameterized family of graphs to generate candidate partitions and solve PCut on these partitions. The parameterization of graphs is based on adaptively modulating node degrees in varying levels to adapt to different levels of imbalanced data. Our framework automatically selects the parameters based on PCut objective, and can be used in conjunction with other spectral partition methods such as 1-spectral clustering, cheeger cut or sparsest cut. Our idea is then justified through limit cut analysis and demonstrated by both synthetic and real experiments.

6. REFERENCES

[1] H. He and E.A. Garcia, "Learning from imbalanced data," in *IEEE Trans. on Knowledge and Data Engineering*, 2009.

[2] L. Hagen and A. Kahng, "New spectral methods for ratio cut partitioning and clustering.," in *IEEE Trans. Computer-Aided Design, 11(9)*, 1992, pp. 1074–1085.

[3] J. Shi and J. Malik, "Normalized cuts and image segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 8, pp. 888–905, 2000.

[4] X. Zhu, "Semi-supervised learning literature survey," in *Technical Report 1530*. 2006, University of Wisconsin-Madison.

[5] U. von Luxburg, "A tutorial on spectral clustering," *Statistics and Computing*, vol. 17, no. 4, pp. 395–416, 2007.

[6] Giulia Galbiati, "Approximating minimum cut with bounded size," in *INOC'11*, 2011, pp. 210–215.

[7] X. Ji, *Graph Partition Problems with Minimum Size Constraints*, Ph.D. thesis, Rensselaer Polytechnic Institute, 2004.

[8] M. Maier, U. von Luxburg, and M. Hein, "Influence of graph construction on graph-based clustering," in *Neural Information Processing Systems 21*. 2008, pp. 1025–1032, MIT Press.

[9] T. Jebara, J. Wang, and S.F. Chang, "Graph construction and b-matching for semi-supervised learning," in *International Conference on Machine Learning*, 2009.

[10] L. Zelnik-Manor and P. Perona, "Self-tuning spectral clustering," in *Neural Information Processing Systems 17*, 2004.

[11] B. Nadler and M. Galun, "Fundamental limitations of spectral clustering," in *Neural Information Processing Systems 19*. 2006, pp. 1017–1024, MIT Press.

[12] T. Buhler and M. Hein, "Spectral clustering based on the graph p-laplacian," in *International Conference on Machine Learning*, 2009.

[13] T. Shi, M. Belkin, and B. Yu, "Data spectroscopy: Eigenspaces of convolution operators and clustering," in *Ann. Statist.*, 2009.

[14] H.D. Simon and S.H. Teng, "How good is recursive bisection?," in *SIAM J. Sci. Comput.*, 1997.

[15] U. Feige and O. Yahalom, "On the complexity of finding balanced oneway cuts," in *Information Processing Letters*, 2003, vol. 87, pp. 1–5.

[16] K. Andreev and H. Racke, "Balanced graph partitioning," in *ACM SPAA*, 2004.

[17] F. Hoppner and F. Klawonn, "Clustering with size constraints," in *Computational Intelligence Paradigms*, 2008.

[18] S. Zhu, D. Wang, and T. Li, "Data clustering with size constraints," in *J. Knowledge-Based Systems*, 2010.

[19] Y. Kawahara K. Nagano and K. Aihara, "Size-constrained submodular minimization through minimum norm base," in *International Conference on Machine Learning*, 2011.

[20] J. Qian and V. Saligrama, "New statistic in p-value estimation for anomaly detection," in *IEEE SSP*, 2012, pp. 393–396.

[21] M. Hein and T. Buhler, "An inverse power method for nonlinear eigenproblems with applications in 1-spectral clustering and sparse pca," in *Neural Information Processing Systems*, 2010.

[22] A. Szlam and X. Bresson, "Total variation and cheeger cuts," in *International Conference on Machine Learning*, 2010.

[23] S. Rao S. Arora and U. Vazirani, "Expander flows, geometric embeddings and graph partitioning," in *Journal of the ACM*, 2009, vol. 56.

[24] A. Frank and A. Asuncion, "UCI machine learning repository," `http://archive.ics.uci.edu/ml`, 2013.

[25] J. Wang, T. Jebara, and S.F. Chang, "Graph transduction via alternating minimization," in *International Conference on Machine Learning*, 2008.