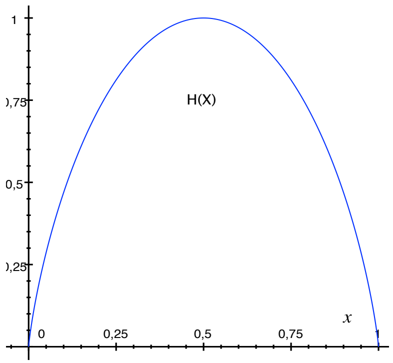
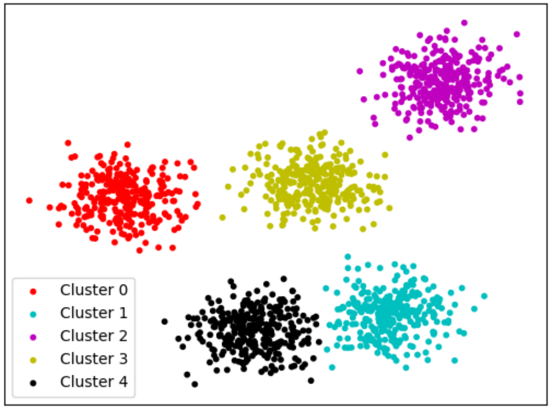




On the Approximability of Information Theoretic Clustering

Ferdiando Cicalese, U. Verona
Eduardo Laber, PUC-RIO
Lucas Murtinho, PUC-RIO

POSTER 165, Pacific Ballroom

Impurity Measures

- Maps a vector **v** in R^d into a non-negative value
- The more homogeneous **v** with respect to its components the larger the impurity
  - (1,0,0,19): small impurity
  - (5,5,5,5) : large impurity

- Well known impurity measures

$$I_{Ent}(\mathbf{v}) = \|\mathbf{v}\|_1 \sum_{i=1}^{g} \frac{v_i}{\|\mathbf{v}\|_1} \log \frac{\|\mathbf{v}\|_1}{v_i},$$

Entropy

$$I_{Gini}(\mathbf{v}) = \|\mathbf{v}\|_1 \sum_{i=1}^{g} \frac{v_i}{\|\mathbf{v}\|_1} \left(1 - \frac{v_i}{\|\mathbf{v}\|_1}\right)$$

Gini

Clustering with minimum impurity

Input

- ***V*** : set of non-negative vectors in R^d
- ***I*** : impurity measure
- ***k*** : number of clusters

Goal

Partition *V* into *k* groups $\mathcal{P} = (V^{(1)}, \ldots, V^{(k)})$ *so that*

$$I(\mathcal{P}) = \sum_{i=1}^{k} I(V^{(i)})$$

is minimized

$I(V^{(i)})$: impurity of the sum of the vectors in $V^{(i)}$

Applications/ Motivations

- Generalizes clustering using KL-divergence
  - Entropy impurity and KL-divergence of a clustering differ by a constant factor
- Clustering probability distributions
- Clustering nominal attributes in decision tree/ random forest construction
- Channel Quantizer Design [Inf. Theory]

Our Contributions

Approximation Algorithms

- 3-approximation for Gini in linear time (arbitrary k)

- $O(\log^2(\min\{d,k\}))$-approximation for Entropy in polytime
  - First algorithm with approximation independent of *n* that does make assumption on the input domain

Our Contributions

Approximation Algorithms

Project vectors in dimension k
incur small additive loss

- 3-approximation for Gini in linear time (arbitrary k)
- $O(\log^2(\min\{d,k\}))$-approximation for Entropy in polytime
  - First algorithm with approximation independent of *n* that does make assumption on the input domain

Our Contributions

Approximation Algorithms

Project vectors in dimension k
incur small additive loss

- 3-approximation for Gini in linear time (arbitrary k)

Each cluster is ***pure***: all vectors have the same largest component

- $O(\log^2(\min\{d,k\}))$-approximation for Entropy in polytime
  - First algorithm with approximation independent of *n* that does make assumption on the input domain

Our Contributions

Approximation Algorithms

Project vectors in dimension k
incur small additive loss

- 3-approximation for Gini in linear time (arbitrary k)

Each cluster is ***pure***: all vectors have the same largest component

- $O(\log^2(\min\{d,k\}))$-approximation for Entropy in polytime

There is a clustering with ***exactly one non-pure*** cluster and impurity $O(\log^2 d) \cdot OPT$
Find this clustering in a 2-dim projection using DP

Our Contributions

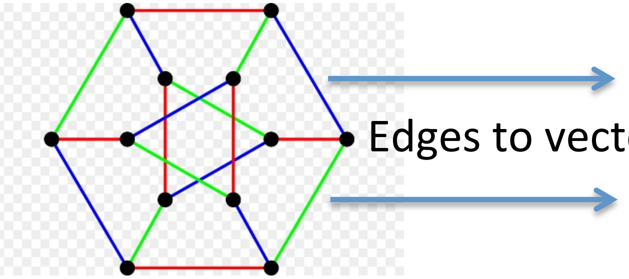
APX-Hardness for Entropy

- Reduction from c-gap vertex cover in cubic graphs

- Solves open question from [Chaudhuri and McGregor, COLT08] and [Ackermann et al., ECCC11]

Our Contributions

APX-Hardness for Entropy



- Reduction from c-gap vertex cover in cubic graphs

Theorem

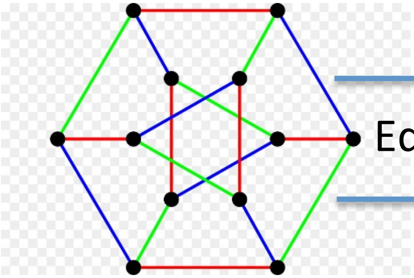
k'(G,k) = 3 log 3|E|+6(1-log3)k

- MinVertexCover ≤ k ⇒ Opt-Impurity ≤ k'(G,k)
- MinVertexCover > ck ⇒ Opt-Impurity > c'k'(G,k)

- Solves open question from [Chaudhuri and McGregor, COLT08] and [Ackermann et al., ECCC11]

Our Contributions

APX-Hardness for Entropy



0..010...010...00

Edges to vectors with two 1's

0..000...010...01

- Reduction from c-gap vertex cover in cubic graphs

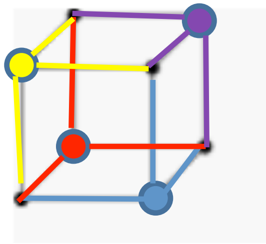
Theorem

k'(G,k) = 3 log 3|E|+6(1-log3)k

- MinVertexCover ≤ k ⇒ Opt-Impurity ≤ k'(G,k)
- MinVertexCover > ck ⇒ Opt-Impurity > c'k'(G,k)

- Solves open question from [Chaudhuri and McGregor, COLT08] and [Ackermann et al., ECCC11]

Lemma. *G cubic and min-VertexCover <= k ⇒ G decomposes into stars of sizes 2 and 3.*



Our Contributions

Ratio-Greedy Algorithm

- Built on top of the theoretical ideas
- Promising preliminary experimental comparisons
  - much faster than a k-means based method
  - close impurity

Our Contributions

Ratio-Greedy Algorithm

- Built on top of the theoretical ideas
- Promising preliminary experimental comparisons
  - much faster than a k-means based method
  - close impurity

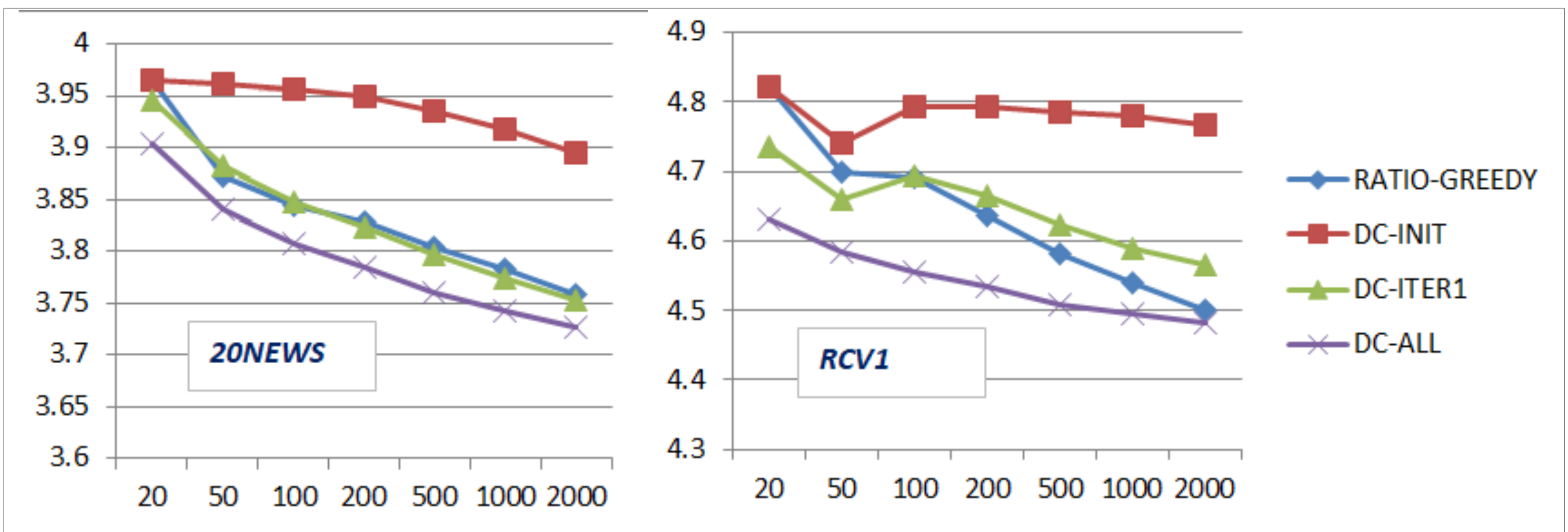
Clusters	RATIO-GREEDY	DC-INIT	DC-ITER1	DC-ITER5	DC-ALL
20	0.6	0.4	3	11	68.6
50	1	0.4	6	25.3	342.1
100	3.1	0.5	10.8	49.1	971.1
200	3.1	0.5	20.3	96.7	1932.4
500	3.3	0.5	48.8	238.7	4823.4
1000	3.5	0.5	96.6	477.2	9612.6
2000	3.5	0.5	191.3	958	19320.2

Our Contributions

Ratio-Greedy Algorithm

- Built on top of the theoretical ideas
- Promising preliminary experimental comparisons
  - much faster than a k-means based method
  - close impurity

Clusters	RATIO-GREEDY	DC-INIT	DC-ITER1	DC-ITER5	DC-ALL
20	0.6	0.4	3	11	68.6
50	1	0.4	6	25.3	342.1
100	3.1	0.5	10.8	49.1	971.1
200	3.1	0.5	20.3	96.7	1932.4
500	3.3	0.5	48.8	238.7	4823.4
1000	3.5	0.5	96.6	477.2	9612.6
2000	3.5	0.5	191.3	958	19320.2



New Results on Information Theoretic Clustering

Ferdinando Cicalese[a] & Eduardo Laber[b] & Lucas Murtinho[b]

[a]Department of Computer Science, University of Verona [b]Departamento de Informática, PUC-RIO

UNIVERSITÀ di VERONA Dipartimento di INFORMATICA PUC RIO DEPARTAMENTO DE INFORMÁTICA PUC-RIO

ICML | 2019

Thirty-sixth International Conference on Machine Learning

Abstract

We study the problem of optimizing the clustering of a set of vectors when the quality of the clustering is measured by the Entropy impurity measure. This is typical of situations where items to be clustered are represented by vectors of frequency counts or probability distributions. Our results contribute to the state of the art both in terms of best known approximation guarantees and inapproximability bounds.

Problem Definition

An impurity measure $I : \mathbf{v} \in \mathbb{R}^d \mapsto I(\mathbf{v}) \in \mathbb{R}^+$ is a function that assigns a vector $\mathbf{v}$ to a non-negative value $I(\mathbf{v})$ so that the more homogeneous $\mathbf{v}$, with respect to the values of its coordinates, the larger its impurity. A well-known example of impurity measure is the Entropy impurity (aka Information Gain in the context of random forests):

$$I_{Ent}(\mathbf{v}) = \|\mathbf{v}\|_1 \sum_{i=1}^{d} \frac{v_i}{\|\mathbf{v}\|_1} \log \frac{\|\mathbf{v}\|_1}{v_i}.$$

Given a collection of n many d-dimensional vectors V with non-

See you tonight!

Pacific Ballroom #165

sulting optimization criterion is closely related (and in some cases equivalent) to minimizing the Entropy impurity.

- **Quantization of discrete memoryless channels.** In this case, the goal is to build quantizations that maximizes the mutual information between channel input and quantizer's output. This is also directly expressible as an instance of PMWIP$_{Ent}$ [11, 15].
- **Attribute selection for decision trees/random forests.** The partition of the values of the attributes during the branching phase in the construction of the decision tree is done by optimizing the change in impurity due to the split [4, 8].

Our Contributions

- a simple linear time algorithm that guarantees
  (i) $O(\log \sum_{\mathbf{v} \in V} \|\mathbf{v}\|_1)$ approximation for PMWIP$_{Ent}$;
  (ii) $O(\log n + \log d)$ approximation for the case where all vectors in V have the same ℓ_1 norm.
- a second algorithm providing $O(\log^2(\min\{k, d\}))$-approximation for PMWIP$_{Ent}$ in polynomial time. This is the first algorithm for clustering based on entropy minimization, that guarantees approximation and does not depends on n.
- an inapproximability result showing that PMWIP$_{Ent}$ is APX-hard even for the case where all vectors have the same ℓ_1-norm. This result solves a problem that remained open in previous work [6, 2].
- some experimental evaluation of a new clustering method developed on top of our theoretical tools/findings with the aim of assessing their potential in practical applications.

Related Work

- **Theoretical results on the structure of the optimal solution.** The PMWIP$_{Ent}$ can be solved in polynomial time when $d = 2$ [11]. This is based on a characterization of the optimal partition in terms of hyperplanes in R^d [7, 5, 8], which provides an $O(n^d)$ optimal algorithm for $k = 2$. For unbounded dimension d, the PMWIP$_{Ent}$ is NP-hard even for $k = 2$. For $k = 2$, constant approximation algorithms have been given for a class of impurity measures including I_{Ent} [13]. These algorithms do not extend to $k > 2$.
- **Clustering probability distributions.** PMWIP$_{Ent}$ is a generalization of MTC$_{KL}$ [6], the problem of clustering a set of n prob-

dimensionality reduction.

Dom(V, k)
1: If $d < k$ create $k - d$ new components for each vector, all of them with 0's
2: Reorder components of all vectors so that for $\mathbf{u} = \sum_{\mathbf{v} \in V} \mathbf{v}$ it holds that $u_i \geq u_{i+1}$ for $i = 1, \ldots, d-1$
3: Let $\mathbf{e}_i$ be the ith standard direction, $i < k$, and $\mathbf{e}_k = 1 - \sum_{i=1}^{k-1} \mathbf{e}_i$
4: Project each $\mathbf{v} \in V$ into $Span(\{\mathbf{e}_1, \cdots, \mathbf{e}_k\})$
5: $V_i \leftarrow \{\mathbf{v} \mid$ largest component of $proj(\mathbf{v}) = i\}$
6: **return** the partition $(V_1, \ldots, V_k)$

We have the following result regarding algorithm Dom.

Theorem. Dom *is a linear time* $O(\log(\sum_{\mathbf{v} \in V} \|\mathbf{v}\|_1))$*-approximation algorithm for* PMWIP$_{Ent}$.

Remark. Dom *also guarantees 3-approximation when the Gini impurity measure is used instead of* I_{Ent}. *This result is tight in the sense that Gini minimization is APX-hard [12].*

$O(\log^2 \min\{d, k\})$- approx for PMWIP$_{Ent}$

- The first step of the algorithm is to employ an extension of the approach introduced in [13] to reduce the dimension of the vectors in V to k, if $d > k$. This step incurs an $O(\log k)$ additive loss in the approximation ratio.
- The remaining steps are based on the following results:
  (i) the existence of an optimal algorithm for $d = 2$ [11];
  (ii) the existence of a mapping $\chi : \mathbb{R}^d \mapsto \mathbb{R}^2$ such that for a set of vectors B which is pure, i.e., a set of vectors with the same dominant component, $I_{Ent}(\sum_{\mathbf{v} \in B} \mathbf{v}) = O(\log d) I_{Ent}(\sum_{\mathbf{v} \in B} \chi(\mathbf{v}))$;
  (iii) a structural theorem that states that there exists a partition whose impurity is at an $O(\log^2 d)$ factor from the optimal one and such that at most one of its groups is mixed, i.e., it is not pure.

A partition of this type with low impurity is constructed using Dynamic Programming over the vectors obtained via the mapping χ – this yields a pseudo-polynomial time complexity. To obtain a polynomial time algorithm, a filtering technique similar to that used in the FPTAS for the subset sum problem is employed.

Inapproximability results

to MTC$_{KL}$. Then, we have

Theorem. *The* PMWIP$_{Ent}$ *is APX-Hard even for the case where all vector have the same* ℓ_1 *norm. Hence,* MTC$_{KL}$ *is APX-hard.*

Experiments

Although the focus of our research is mainly theoretical, we also designed Ratio-Greedy, a fast and practical algorithm that relies on our theoretical results.

Ratio-Greedy(V, k)
1: **if** $k \leq d$ **then return** Dom(V, k)
2: **Divide** V into d sets $V_1, \ldots V_d$, according to the largest component
3: **Sort** each V_i into a list L_i of singleton clusters $\{\mathbf{v}\}$ sorted according to $ratio(\mathbf{v}) = \|\mathbf{v}\|_1 / (\|\mathbf{v}\|_1 - \|\mathbf{v}\|_\infty)$
4: **Reduce** the number of clusters from n to k by applying the following operations:
5: **Pick** a pair C, C' of adjacent clusters in some L_i that minimizes $loss(C, C') = I_{Ent}(C \cup C') - I_{Ent}(C) - I_{Ent}(C')$
6: **Replace** C, C' with $C \cup C'$
7: **return** the collection of resulting clusters in the d lists

Complexity and guarantee

- Ratio-Greedy can be implemented to run in $O(n \log n + nd)$ time, exploiting a binary heap to select the adjacent clusters in L_i whose merge incurs the minimum $loss$.
- The impurity of the partition obtained by Ratio-Greedy is no worse than that obtained by Dom due to the superadditivity of I_{Ent}, thus it inherits its approximation guarantees.

Baseline. We compared Ratio-Greedy with Divisive Clustering (DC for short), an adaptation of the k-means method proposed in [9] to solve PMWIP$_{Ent}$.

Datasets. We tested these methods on clustering 51.480 words from the 20NEWS corpus and 170.946 words from RCV1 corpus, according to their distributions w.r.t. 20 and 103 different classes respectively.

Result analysis. The figure below shows the impurities of the partitions obtained for different values of k for both datasets. DC-Init, DC-Iter1 and DC-All correspond, respectively, to different points in the execution of DC: right after its initialization, after its first iteration and at the end.

[3] M.R. Ackermann, J. Blömer, C. Sohler. Clustering for metric and nonmetric distance measures. *ACM Trans. Algorithms*, 6(4):59:1–59:26, 2010.
[4] L. Breiman, J.J. Friedman, R.A. Olshen, C.J. Stone. *Classification and Regression Trees.* Wadsworth, 1984.
[5] D. Burshtein, V. Pietra, D. Kanevsky, A. Nadas. Minimum impurity partitions. *Ann. Stat.*, 1992.
[6] K. Chaudhuri, A. McGregor. Finding metric structure in information theoretic clustering. In *Proc. of COLT 2008*, pp. 391–402. 2008.
[7] P. A. Chou. Optimal partitioning for classification and regression trees. *IEEE Trans. on Pattern Analysis and Mach. Int.*, 13(4), 1991.
[8] D. Coppersmith, S.J. Hong, J.R.M. Hosking. Partitioning nominal attributes in decision trees. *Data Min. Knowl. Discov*, 3(2):197–217, 1999.
[9] I.S. Dhillon, S. Mallela, R. Kumar. A divisive information-theoretic feature clustering algorithm for text classification. *JMLR*, 3:1265–1287, 2003.
[10] S. Jegelka, S. Sra, and A. Banerjee. Approximation algorithms for Bregman co-clustering and tensor clustering. *CoRR*, abs/0812.0389, 2008.
[11] B.M. Kurkoski, H. Yagi. Quantization of binary-input discrete memoryless channels. *IEEE Trans. Inf. Th.*, 60(8):4544–4552, 2014.
[12] E. S. Laber, L. Murtinho. Minimization of gini impurity: NP-completeness and approximation algorithms via connections with the k-means problem. In *Proc. of LAGOS*, 2019. To appear.
[13] E. Laber, M. Molinaro, F. Mello. Binary Partitions with Approximate Minimum Impurity. In *Proc. of ICML 2018*, vol. 80 of *Proc. MLR*, pp. 2854–2862, 2018.
[14] M. Lucic, O. Bachem, A. Krause. Strong coresets for hard and soft Bregman clustering with applications to exponential family mixtures. In *Proc. of Machine Learning Res.* , pp. 1–9, 2016.
[15] U. Pereg, I. Tal. Channel upgradation for non-binary input alphabets and MACs. *IEEE Trans. Inf. Th.* , 63(3):1410–1424, 2017.