M-021

近似 k 近傍グラフの作成による計算コスト削減手法の提案

A Proposed Method to Reduce Computational Costs for Creating Approximate k Neighbor Graph

廣中 雅大 † 後藤 佑介 †
Masahiro Hironaka Yusuke Gotoh

1. はじめに

近年，k 近傍グラフはデータベースの索引付けや画像検索に利用され，大きな注目を集めている．しかし，探索対象の事例数が増加すると k 近傍グラフの作成にかかる計算量は膨大となり，処理時間は長大化する．

本研究では，提案手法を用いて近似 k 近傍グラフを作成することで，k 近傍グラフの作成に比べて計算コストを削減する手法を提案する．提案手法では，類似探索手法の一つである Locality Sensitive Hashing (LSH)，および近似 k 近傍グラフ構築手法の一つである NN-Descent 法を用いて近似 k 近傍グラフを作成する．また，このグラフに半教師あり学習を適用することで，分類精度を維持しつつ，近似 k 近傍グラフの作成および探索にかかる時間を短縮する．評価では，k 近傍グラフと比較して，提案手法の有用性を確認する．

2. 最近傍探索

2.1. 概要

最近傍探索では，問い合わせ元となる事例 (以下，クエリ) からもっとも距離が近い事例を探索する．単純な最近傍探索である線形探索では，クエリとすべての事例との間の距離を計算するため，探索にかかる計算量は $O(n)$ となる．また，k 近傍探索では，クエリからの距離が近い順番に k 個の事例を探索する．

2.2. 近似最近傍探索

近似最近傍探索では，クエリからもっとも近い事例を厳密に計算せず，近似した事例を探索する．近似最近傍探索による探索結果は最近傍の厳密な解ではないが，最近傍探索と比較して高速に探索できる．近似最近傍探索を行う手法として，Locallity Sensitive Hashing(LSH)[1][2] がある．

2.3. 球面集中現象

高次元のデータに最近傍探索を行う場合，クエリとクエリから一定距離内に存在する各オブジェクトとの距離の差が低次元の最近傍探索と比較して小さくなる球面集中現象が発生する．このとき，最近傍点，および近似最近傍点の探索に必要な計算量が増加する．このため，ユークリッド空間を階層的に空間分割する最近傍探索手法では，高次元空間において線形探索と同様の計算量が必要になる．

3.LSH

3.1. 概要

LSH は，ハッシュを用いた近似最近傍探索手法である．類似したデータは同じハッシュ値をもつ性質を利用して，2 点の距離が近いほどハッシュ値が同じ値をもつ確率が高くなるハッシュ関数を用いることで，最近傍探索の近似解を求めて探索を高速化する．

3.2.LSH を用いた近似最近傍探索

LSH を用いた近似最近傍探索では，複数個のハッシュ関数を用いることで，近似最近傍の候補となる点の精度を向上させる．また，ハッシュ関数が一個となる場合，k 近傍となる点が近似 k 近傍の候補点にならない可能性がある．そこで，ハッシュ関数を L 個作成し，L 個のハッシュ関数に対してハッシュテーブルを N 個作成し，クエリとハッシュ値が一致する点を登録する．ハッシュテーブル内に存在する点を用いて線形探索を行い，k 近傍点を決定する．テーブル数 N が多いほど計算量は増えるが，真の最近傍点を探索できる確率が高くなる．

4. 近傍グラフ

4.1.k 近傍グラフ

k 近傍グラフ [3] は，各事例に対して k 近傍となる事例との間に辺を作ることで作成するグラフである．このとき，事例間の辺は有向辺と無向辺の 2 種類があり，本研究では無向辺を用いる．

d 次元空間上に存在する事例を $x_1, x_2, \cdots, x_n$ とし，各事例から構築したグラフを G とする．G の辺は n 行 n 列の行列で表す．

4.2. 近似 k 近傍グラフ

k 近傍グラフの計算量は $O(dn^2)$ となるため，大規模なデータをもとに k 近傍グラフを構築する場合，膨大な時間が必要となる．このため，k 近傍グラフの構築にかかる計算量を削減する近似 k 近傍グラフが提案されている．近似 k 近傍グラフ内の任意の事例に対する k 近傍点は，必ずしも真の k 近傍点であるとは限らない．

4.3.NN-Descent 法

NN-Descent 法では，はじめに事例間でランダムに辺を追加したグラフ (以下，ランダムグラフ) を作成する．ランダムグラフの更新では，更新の対象となる頂点は，自身から隣接する頂点に隣接する頂点までの範囲に限定して探索することで，計算量を削減する．

5. 半教師あり学習

5.1. 半教師あり学習の定義

d 次元空間上での n 個の事例に対する集合を $X = x_1, \cdots, x_n$，n 個の事例に対するラベルの集合を $Y = y_1, \cdots, y_n$ とする．ここで，$y_1, \cdots, y_l$ $(1 \leq l \leq n)$ はラベルが既知であり，残りの $y_{l+1}, \cdots, y_n$ は不明である．半教師あり学習では，ラベルが既知である $y_1, \cdots, y_l$ を用いて，ラベルが不明である $y_{l+1}, \cdots, y_n$ を予測する．

†岡山大学大学院自然科学研究科

5.2. グラフ構造を用いる手法

グラフ構造を用いた半教師あり学習は，グラフ構築とラベル伝搬の 2 段階に分類できる．グラフ構築の段階では，すべての事例を頂点とし，各事例間の距離をもとにグラフを構築する．ラベル伝搬の段階では，構築したグラフをもとに，ラベルデータから未ラベルデータに伝搬したラベルを付与する．本研究では，グラフ構造を用いた半教師あり学習でラベルを伝搬する手法として，Local and Global Consistency(LGC)[4] を用いる．

6. 提案手法

6.1. 概要

LSH を用いた近似 k 近傍グラフの作成手法を説明する．提案手法では，各事象に対して LSH を用いた近似最近傍探索を行い，この探索結果を用いて作成した近似 k 近傍グラフに対して，NN-Descent 法を用いてグラフを更新する．LSH を用いた近似最近傍探索は，球面集中現象による計算量の増加を抑えることができる．また，作成した近似 k 近傍グラフを用いた分類精度を向上させるため，NN-Descent 法を用いる．これにより，分類精度を維持しつつグラフ構築を高速化できる．

6.2. 処理手順

提案手法の処理手順は以下の通りである．

1. 各点に対して LSH を用いた近似 k 近傍探索を行い，各ハッシュテーブルで k 近傍点の候補を求める．
2. 各点と自身の k 近傍点との間に無向辺を作成する．
3. 2. で作成した k 近傍グラフに対して NN-Descent 法を用いて更新し，k 近傍点を決定する．

7. 評価

表 1 に，MNIST のデータセットを用いた分類精度を示す．MNIST は，0 から 9 までの数字を手書きで書かれた画像データの集合であり，データの大きさは 28 × 28 ピクセル，画像は 70,000 枚である．これらのデータをクラスラベルが 1 から 9 の 10 種類に分類する．データの構成として，全体の 0.5%をラベルデータ，および残りの 95.5%を未ラベルデータとする．

表 1 より，提案手法では，テーブル数が増加するとハッシュ値が同じとなる事例数が増加し，距離計算の回数が増加するため，処理時間は長大化する．また，k 近傍点の候補となる事例数が増加して，真の k 近傍点を選択する可能性が高くなるため，分類精度は向上する．

次に，テーブル数が 50 の場合，提案手法における処理時間は，k 近傍グラフと比較して約 48.1%短縮した．提案手法では，LSH で k 近傍点の近似解を求めることで，距離計算の回数を減少できる．また，提案手法における分類精度は，k 近傍グラフと比較して差は約 0.002 とわずかであった．提案手法では，NN-Descent 法を用いることで，より k 近傍グラフに近い近似 k 近傍グラフを作成でき，分類精度が向上する．

表 1: MNIST を用いた処理時間および分類精度

手法	テーブル数	処理時間 (秒)	分類精度
k 近傍グラフ	-	8,860	0.87997
提案手法	10	1,467	0.83381
	20	2,166	0.86137
	30	3,072	0.86283
	40	3,957	0.86632
	50	4,593	0.87790

8. まとめ

本研究では，グラフ構造を用いる半教師あり学習において，LSH を用いた近似最近傍探索と近似 k 近傍グラフの作成，および NN-Descent 法によるグラフ構築を行い，近似 k 近傍グラフの構築にかかる計算量を削減する手法を提案した．提案手法では，LSH を用いた近似最近傍探索を行うことで，k 近傍グラフの作成に比べて計算コストを削減し，NN-Descent 法を用いて分類精度を維持する．MNIST を用いた評価において，テーブル数が 50 の場合，提案手法を用いて作成する近似 k 近傍グラフは，k 近傍グラフと比較してグラフの構築にかかる処理時間を約 48.1%短縮するとともに，分類精度の低下は約 0.002 とわずかであった．

謝辞

本研究の一部は，文部科学省科学研究費補助金・基盤 (B)(15H02702)，基盤 (C)(16K01065)，および (公財) 中島記念国際交流財団の研究助成による成果である．

参考文献

[1] P. Indyk and R. Motwani: Approximate Nearest Neighbors: Towards Removing the Curse of Dimensionality, Proc. 30th ACM Symposium on Theory of Computing, pp.604-613 (1998)

[2] M. Datar, N. Immorlica, P. Indyk, and V. S. Mirrokni: Locality-Sensitive Hashing Scheme Based on p-stable Distributions, Proc. 21st ACM symposium on Computational Geometry, pp.253-262 (2004)

[3] T. Sebastian and B. Kimia: Metric-based Shape Retrieval in Large Databases, Proc. 16th International Conference on Pattern Recognition, Vol.3, pp.291-296 (2002)

[4] D. Zhou, O. Bousquet, T. Navin, J. Weston, and B. Scholkopf: Learning with Local and Global Consistency, Proc. Advances in Neural Information Processing Systems, pp.321-328 (2004)