F-017

社会インフラシステムの稼働データ流通のための バイナリハッシュを用いた近似クラスタリング
Approximate Clustering with Binary Hashing for Operational Data Distribution in Public Infrastructure

但馬慶行† 望月智之† 志村 明俊† 武澤 隆之‡
Yoshiyuki Tajima Tomoyuki Mochizuki Akitoshi Shimura Takayuki Takezawa

1. はじめに

鉄道システムや電力システムといった社会インフラシステムは、制御を含む広域に分散した複雑システムである。近年、老朽化したシステムのメンテナンス費用を削減するため、異常検知や予防保守などの高度なサービスにも期待が集まっている[1]。こうしたサービスにとって、設備や機器が生成する稼動データは粒度が細かい。そのため、目的に応じて稼動データを適宜要約しながら流通する必要がある。

本稿では、まず異常検知を例とした稼動データ流通について簡単に説明する。次に異常検知とクラスタリングの関係性を述べ、社会インフラシステムの特性上、クラスタリングの高速化が課題となることを述べる。そして、クラスタリングにおいて比較的負荷のかかる距離の計算を近似することで高速化する手法を提案する。

2. 稼動データ流通とその課題

本稿の前提となる典型的な社会インフラシステムの構成と、異常検知における稼動データの流通を図 1 に基づいて説明する。典型的な社会インフラシステムは、広域に分散したモノの制御を担う制御システム、それらを統括する中央指令システム、保守管理システムなどの情報システムから構成される。各制御システムの設備や機器、それらを制御する制御サーバが稼動データを生成する。生成された稼動データは、専用回線等を介して中央指令システムに送られる。この稼動データにはユーザの利用実績など個人情報を含む場合がある。そこで中央指令システムは、稼動データを使って異常検知を行うことで要約し、その要約結果のみを保守管理システムに送る。このように、社会インフラシステムでは、しばしば稼動データを用途や制約に応じて要約しながら流通させる。

次に異常検知の流れとクラスタリングの関係を説明する。異常検知の一つの考え方は、正常時の稼動データから期待される稼動データの生成モデルを学習しておいて、そのモデルと新しく得た稼動データの乖離に基づき異常を検知することである。ここで、制御システムには、複数の動作パターンがあり、その動作パターンごとに稼動データの分布が大きく異なる。このため、例えば、稼動データの発生周期や動作順序を特徴量としてユークリッド距離などに基づくクラスタリングを用いてクラスタを分割し、新しく得たデータと最も近いクラスタの近傍点からの距離で乖離を評価する、といった工夫が必要となる。その際、稼動データが大量になると、一般的にクラスタリングの計算量は多大なものとなる。

†株式会社 日立製作所 横浜研究所
‡株式会社 日立製作所 インフラシステム社

次に運営形態に伴う社会インフラシステムの特性と本研究の課題を述べる。社会インフラシステムは何十年にも渡って段階的な拡張がなされる。この結果、制御システムが生成する総稼動データ量は年々増大する。これに伴い、稼動データを捌くための計算機リソースを増強する必要がある。これに対し保守管理システムなどの情報システムは、コンセッション契約を結んだ企業がクラウド上に構築する場合も増えており、容易に増強できる場合が多い。一方、制御の中核を担う図 1 の中央指令システムは、信頼性や安全性の観点から公共性の高い事業者がオンプレミスで構築している。このため、保守を担う企業の一存で増強もできないし、増強するとしても多大な時間と費用を要する。したがって、今ある計算機リソースである程度の拡張に対応できることのメリットは大きい。そこで、本研究では制御系で計算負荷の大きいクラスタリングを高速化することを課題とする。

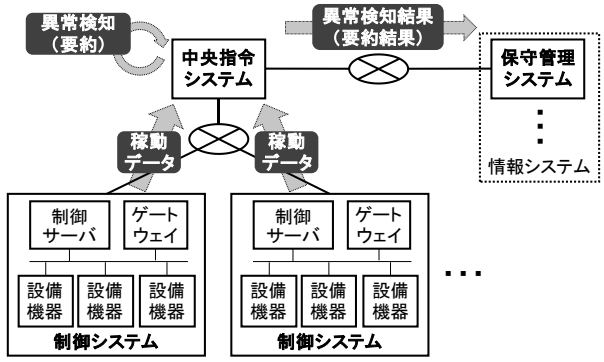

図 1 社会インフラシステムの構成と稼動データの流通

3. 近似クラスタリング手法の提案

ある 2 つのデータの距離が大きい場合、その距離を近似しても実用的な結果が得られるという仮説のもと、クラスタリングにおける距離計算を近似することで高速化する手法を提案する。

3.1 バイナリハッシュによる近似クラスタリング

提案手法の基本方針は、クラスタリングの距離計算について、軽量な計算で距離が大きいかどうかを判別し、大きい場合には近似値に置き換えるというものである。この距離が大きいかどうかを判別に、元の空間の距離関係を保存し、かつ、距離計算が高速に実行できるバイナリハッシュ[2]を用いる。

具体的な手順は次のとおりである。稼動データ全体Uのある稼動データ$\mathrm{X_i}, \mathrm{X_j} \in U$に対するハミング距離を$\mathrm{H(i,j)}$

とする。このとき、H(i, j)が閾値αより小さければ本来の距離$d(i,j)$を計算し、それ以外は、計算量の小さい距離d'の近似値に置き換える。閾値αは稼動データの重要度や関係性に応じて変えることもよい。この近似処理を行った距離$distance(i,j)$を式(1)に示す。

$$\text{distance(i,j)} = \begin{cases} d(i,j)\ if\ H(i,j) < \alpha \\ d'(i,j)\ otherwise \end{cases} \quad (1)$$

他の稼動データに依存しないである稼動データのハッシュが生成できるハッシュ生成手法を用いることで、稼動データを収集した時点で逐次的にハッシュを生成できる。したがって、ハッシュ生成にかかる計算負荷は制御システムのゲートウェイ等に回すことができる。

3.2 距離の近似値の計算方法

本稿では 2 つの近似方法を述べる。1 つ目は、閾値α以上の場合、すべて 0 に置き換える方法である。本稿ではこれを「0 置換」と呼ぶ。この方法は、元の距離との誤差が大きくなる場合もあるが、結果がスパースとなるため、計算効率の向上、メモリ使用量の削減が可能となる。

2 つ目は、サンプリングした稼動データを使って、ハミング距離に対する元の空間の距離の回帰モデルを構築し、距離を推定する方法である。本稿ではこれを「回帰置換」と呼ぶ。この方法は学習コストがかかるものの 0 置換と比べ誤差を小さくできる。なお、ハミング距離のパターンがハッシュのビット数+1 個なので、高々1 度の計算とメモリの参照で実行できる。

4. 評価実験

本手法をスペクトラルクラスタリング[3]に適用して実験を行った。以下にその内容を示す。

4.1 実験内容

手書きの数字に関するラベルつきの公開データ(訓練用から先頭 5000 件)[4]をスペクトラルクラスタリングで分類する。スペクトラルクラスタリングは、ラプラシアン行列Lの固有ベクトルを計算し、その後固有ベクトルから得られる各要素の特徴ベクトルに対して k-means を実行する。この処理の中で各データの類似度に基づく隣接行列Aが計算される。その際、隣接行列Aの各要素を前章の考え方に従い置き換える。ここで類似度は$1/(1+d(i,j))$($d(i,j)$はユークリッド距離)とした。

バイナリハッシュは M.Raginsky らが提案した手法[2]を用いて生成する。ハッシュサイズは 128bit で固定し、ガウシアンカーネル(γ=0.5)を選択した。$d'(i,j)$は前記の 0 置換と回帰置換を用いる。回帰置換で用いる回帰モデルは 2 次の線形多項式とし、500 件ランダムサンプリングした$d(i,j)$を使って学習する。また、クラスタ数は正解のラベル数と同じ 10 とする。クラスタの初期値は各ラベルからランダムで 1 つずつ選択する。

4.2 実験結果

2 つの近似方法について閾値αを変えて評価した。評価指標として次の指標を設定した。削減率 D は全体の距離計算回数に対する近似した回数の割合を表す。正解率 P は元のデータに付与されている正解のラベルと、クラスタリング結果の純度(Purity)を表す。この値が 1 に近いほど正解のラベルに近い。一致率 Q は近似せずに得られたクラスタリング結果を正解ラベルとしたときの近似した場合のクラスタリング結果の純度を表す。この値が 1 に近いほど近似の影響が小さい。

実験の結果を表 1 に示す。表の値は各αに関し 10 回実行した結果の平均である。このデータとアルゴリズムの組み合わせでは、正解率 P の上限は 0.7 を少し超える程度となっている。2 つの近似方法ともに、閾値が下がるとともに削減率 D が増加し、正解率 P、一致率 Q が減少する傾向がある。0 置換は近似無しの場合に比べ 0.1 近く正解率 P が低下している。回帰置換は半分近く削減しても近似なしの場合と同程度の正解率 P を維持している。

表 1 スペクトラルクラスタリングに適用した実験の結果

近似方法	閾値α	削減率 D	正解率 P	一致率 Q
近似なし	—	0%	0.733	1
0 置換	25	56.1%	0.646	0.695
	18	84.1%	0.628	0.644
	12	94.8%	0.645	0.634
回帰置換	25	48.4%	0.733	0.861
	18	82.4%	0.726	0.814
	12	93.6%	0.723	0.790
	0	100%	0.694	0.792

4.3 考察

スペクトラルクラスタリングでは、全データの関係性を考慮するため、局所的に距離に誤差が大きくなっても、平均的に誤差が小さくなり提案手法が機能したと考えられる。

5. おわりに

本稿ではバイナリハッシュを用いた距離計算の近似による高速化手法を提案した。また半分程度距離計算を近似しても性能が維持できることを実験的に確認した。

今後の課題は、現実問題に即したデータによる評価である。また、実際には固有ベクトル計算などの処理もボトルネックとなるため、次元圧縮[5]などとの組み合わせや、別のクラスタリング手法への適用検討を進める。

参考文献

[1] “各府省庁のインフラ老朽化対策の状況” http://www.cas.go.jp/jp/seisaku/infra_roukyuuka/dai1/sankou.pdf
[2] M.Raginsky, S. Lazebnik, “Locality-Sensitive Binary Codes from Shift-Invariant Kernels”, Advances in Neural Information Processing Systems (NIPS), pages 1509-1517(2009)
[3] U. von Luxburg, “A Tutorial on Spectral Clustering”, Statistics and Computing 17(4), 2007
[4] “Pen-Based Recognition of Handwritten Digits Data Set”, http://archive.ics.uci.edu/ml/datasets/Pen-Based+Recognition+of+Handwritten+Digits
[5] E. Liberty, “Simple and Deterministic Matrix Sketching”, 19th ACM Conference on Knowledge Discovery and Data mining (SIGKDD), Pages 581-588(2013).