2007 年情報論的学習理論ワークショップ
2007 Workshop on Information-Based Induction Sciences (IBIS2007)
Tokyo, Japan, November 5 - 7, 2007.

近傍保存原理による異常検知

Anomaly Detection with Neighborhood Preservation Principle

井手剛*
Tsuyoshi Idé

Abstract: We consider a task of anomaly diagnosis from multiple sensor data. Our goal is to compute the anomaly score of each sensor under conditions that sensor signals are of highly dynamic and correlated natures, where existing techniques are less useful. We treat a set of sensor signals as a graph stream, and decompose the graph into a set of neighborhood graphs. The anomaly score is then defined based on a novel notion of neighborhood preservation. Applying to a real-world sensor validation task, we demonstrate the utility of our approach.
Keywords: sensor data, anomaly detection, neighborhood graph

1 はじめに

時系列データからの知識発見はデータマイニングにおける主要な研究テーマのひとつである。とりわけ最近では、相関のある複数時系列をグラフストリームと捉え、その観点から解析を行う研究が興味を引いている [5, 9, 10, 11]。その場合のグラフとは、各頂点が個々の時系列、辺の重みは時系列同士の（非）類似度を表す。

実用上、実数値の時系列データに対しては、教師なし学習の枠組みで問題を扱うことが要求されることが多い。典型的な問題設定は変化検出（または異常検知）である [13, 12]。変化検出は普通、何かデータ生成機構がデータの裏にあると考えた時、その機構の変化を捉える問題として定式化される。しかし、複数時系列からの知識発見問題を考えた時、変化検出だけで必ずしも話は終わらないことに気付く。すなわち、変化の存在自体を知った後、どの変数（もしくは自由度）がどれだけその変化に関与したのかを知りたい、というのが実用的に広く現れる要請である。この問題を仮に、**変化解析**と呼んでおく。

この論文では、多変数の時系列データからの変化解析の問題を扱う。直接想定する応用としては、自動車などの動的システムの異常診断というのがある。図 1 を参照されたい。対象とするシステムの正常時に、センサー群を介して時系列データを取っておく。異常の疑いがある時に、やはりデータを取る。センサーとしては、速度や

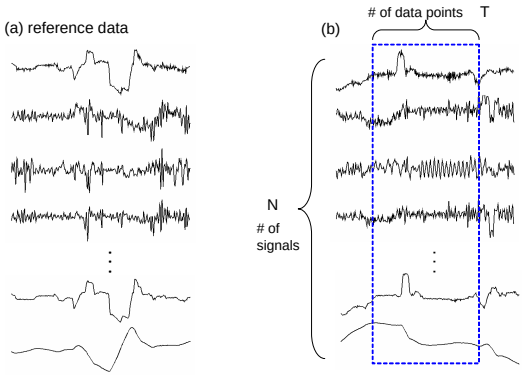

図 1: 問題設定。正常時の時系列データ (a) とテスト時の時系列データ (b) を比べて、各センサーの異常度を計算したい。

加速度など力学的な測定値に由来するものを想定する。「この二組のデータが違うか否か」を述べるのが変化検出の問題であり、「両者がそもそも違うと思ったとき、何がどれだけ効いているのか」を調べるのが変化解析の問題である。言い換えると、図の (a) と (b) のように、2 つの多変数時系列データが与えられた時、各センサーの**異常度**を計算するのがわれわれのタスクである。

この種の問題への解法はデータの特徴に依存している。われわれは、自動車などのセンサーデータを念頭に、次の 4 つの想定を置く。

1. システムは高度に動的である
2. センサの種類が違えば信号の挙動も大きく異なる
3. センサ同士には強い相関が存在する

*IBM 東京基礎研究所, 242-8502 神奈川県大和市下鶴間 1623-14, e-mail: goodidea@jp.ibm.com.
IBM Research, Tokyo Research Labotarory, 1623-14 Shimo-Tsuruma, Yamato-shi, Kanagawa 242-8502, Japan.

4. センサの挙動についての事前知識は与えられない

このうち、1 の意味するところは、図 1 に例示したように、データが異なれば時系列の様相もがらりと変わり、同一のセンサーからの信号であっても「重ねて比べてみる」ことが意味をなさないということである。しかも信号同士には相関があるため、各時系列をばらして何かの特徴抽出にかけることは一般には許されない。

従来、変化解析の問題は、変化検出の副産物として扱われるのが普通であった。たとえば Papadimitriou ら [8] は、複数時系列に逐次的に主成分分析（PCA）を行い、主成分の変化から変化点を求め、その変化への寄与度で各センサーの異常度を計算している。一方 Idé ら [4] はセンサー同士のある種の類似度を入力として、多次元尺度構成法（MDS [1]）で各センサーの「座標」を求め、その座標の変化から異常度を調べる方法を提案している。PCA や MDS は、必然的にデータ全体のグローバルな構造の変化を問題する。しかしながら、少なくとも図 1 に示すような高度に動的な相関系では、データ全体の構造は極めて不安定のはずであり、これらの手法の実用性には相当の限界があると言うべきである。

変化解析は、その定義から、特定のセンサーの周りのローカルな変化に関係している。われわれはこの点に注目し、**近傍性保存原理**とわれわれが呼ぶ原理に従って異常度を計算することを提案する。まずわれわれは、生の時系列の代わりに、それらの（非）類似度を入力として採用し、それをグラフの重み行列とみなす。次いで、確率的 k 近傍 [3, 2] のアイディアを流用して、各センサーの近傍グラフを構成する。これにより、異常度に確率的解釈を与えることができる。また、確率的近傍についての従来の定式化に、「自己結合確率」という概念を導入することにより、他に超然と常にふらふらしているような時系列の寄与を自動的に割り引くことが可能になる。

以下、次節で問題設定を行い、異常度の定義、実験結果、まとめ、と続く。

2 問題設定

この節では、やや形式的に問題を定義しなおし、近傍保存の概念を述べる。

2.1 相関異常解析問題

たとえば圧力や加速度計といった N 個の物理センサーを持つ動的システムを考える。それぞれのセンサーは、N 次元実数値のデータ点 T からなる（図 1 参照）。ひとつのデータ単位は、したがって、$N \times T$ のデータ行列をなすが、これを「ラン」と呼ぶことにする。ひとつのランにおいては、すべてのセンサーは同期して測定がなされるものと仮定する（そうでない場合は補間その他の前処理でそうする）。第 i 番目のセンサー ($i = 1, 2, ..., N$) の、時刻 t ($t = 1, 2, ..., T$) における測定値を $x_i^{(t)}$ と表す。D と $\bar{\mathsf{D}} \in \mathbb{R}^{N \times N}$ をそれぞれ、テスト時と見本データの、センサー同士の非類似度行列とする（非類似度の定義はすぐ後で述べる）。D と $\bar{\mathsf{D}}$ をグラフの重み行列とみなせば、われわれの問題は次のように述べられる。

定義 1 (相関異常解析問題) 重み行列 D を持つテスト・グラフと、同じく $\bar{\mathsf{D}}$ を持つ見本グラフが与えられた時、両者の相違を説明するスコアを、各頂点に対し計算せよ。

以後、見本データについての量を、上棒 (¯) で区別する。行列 D の (i, j) 成分は時系列 i と j の間の非類似度で、これを $d_{i,j}$ と表す。$\bar{\mathsf{D}}$ と $\bar{d}_{i,j}$ の関係も同様である。

われわれは非類似度 $d_{i,j}$ に、以下の条件を要請する。(1) $\forall i$ に対し $d_{i,i} = 0$、(2) 強く相関したペアに対しては $d_{i,i} \approx 0$、(3) ほぼ無相関のペアに対しては $d_{i,i} \to \infty$。ここで第 2 と第 3 の条件を区別することは大切である。時系列同士に何か大きい相関がある場合、それはシステムの内部構造の表れと考えられるので、情報として重視されるべきである。一方、相関が大きくないペアは、仮にその相関が多少変動しても、あまり重視すべきではない。

この要請を満たすような非類似度の選び方には任意性があるが、以下素朴に、$d_{ij} = -\log |a_{i,j}|$ とおく。ただし、数 $\{a_{i,j}\}$ は時系列 i と j の間の相関係数で、共分散

$$c_{i,j} \equiv \frac{1}{T} \sum_{t=1}^{T} [x_i^{(t)} - \langle x_i \rangle][x_j^{(t)} - \langle x_j \rangle]$$

に対して $c_{i,j}/\sqrt{c_{i,i} c_{j,j}}$ で定義される。ただし、$\langle x_i \rangle \equiv \frac{1}{T} \sum_{t=1}^{T} x_i^{(t)}$ である。定数の信号に対しては $a_{i,j} = \delta_{i,j}$ のように決める（$\delta_{i,j}$ はクロネッカーのデルタ）。

2.2 近傍保存原理

あるノード i に対し、k 番目に小さい非類似度を持つノード（i 自身は省く）を、i の第 k 最近傍と定義する。また、あるノード i に対し、その第 1 最近傍から第 k 最近傍までを含む集合を k 最近傍集合と呼び、$\mathcal{N}_i$ と表す。これらを用いて、k 最近傍グラフを次のように定義する。

定義 2 (k 最近傍グラフ) あるノード i の k 最近傍グラフとは、$\mathcal{N}_i$ と i それ自身をノードとして含み、ノード i とその最近傍ノードを辺でつないだものである。

この定義において、i ノードをこの k-近傍グラフの中心ノードと呼ぶ。

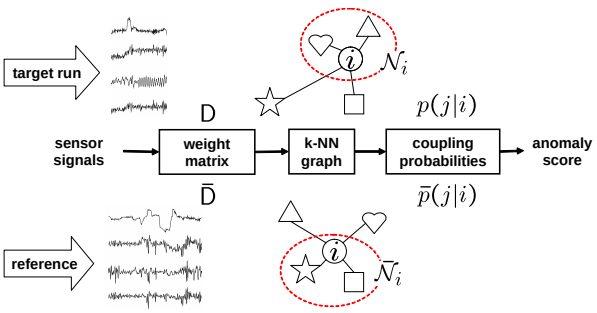

図 2: 全体の流れ。

先に述べたように、自動車や人工衛星といった種類の多センサー系に対するわれわれの前提は、相関の強いセンサー対が存在するということである。そうでない場合、各センサーを独立に考えればよく、グラフストリームとして扱う必要はなくなる。図 1 を見ると、一見このようなデータの変化解析を実行するのは絶望的に思えるが、われわれの観察によれば、センサー間相関におけるデータの揺らぎのほとんどは、ゆるく結合したセンサー対によって担われる。逆に、強く相関したセンサー対については、データが非常に動的であっても、その結合は比較的安定している。これらの事実を「原理」として述べたのが下記である。

定義 3 (近傍保存原理) もしシステムが正常に動作していれば、各ノードの近傍グラフは、実験条件の揺らぎに対してほぼ不変である。

図 2 に、われわれの解析手法をまとめておく。まず、グラフの相違度行列を用いて、各ノードの k-近傍集合を選ぶ。次に、その上で確率分布 $p(j|i)$ が計算される。ただし $p(j|i)$ は、ノード j が、ノード i の k-最近傍グラフに含まれる確率である。

3 確率的 k-近傍グラフ

近傍保存原理は、i ノードを中心ノードとする k-近傍グラフにおける何らかの変化が、i 番目のセンサーの異常と関連付けられることを示唆している。この節では、近傍グラフにおける変化をどう定量的に求めるかについて議論する。なお、本節は、筆者らの別論文 [6] に従う。

3.1 ノード間の結合確率

k 最近傍グラフにおける変化を定量的に表すため、ノード j とノード i の最近傍対となる確率 $p(j|i)$ を導入する。各ノード i に対して、規格化条件は下記のように書かれる。

$$p(i|i) + \sum_{j \in \mathcal{N}_i} p(j|i) = 1 \tag{1}$$

ただし j が i でも $\mathcal{N}_i$ でもない時は、$p(j|i) = 0$ である。項 $p(i|i)$ は、中心ノードが k 最近傍集合のどれとも結合しない確率を表している。

分布 $p(j|i)$ の関数形を決めるため、次のような問題を考える：中心ノード i がその k 最近傍のどれかと、平均して c 本のリンクを取るという条件の下で、近傍グラフをできるだけコンパクトに構成せよ。

手初めに、「k 個の近傍がどれも平等である」と仮定してこの問題の意味するところを考えてみよう。$p(j|i) = \frac{1}{k}(1 - \delta_{i,j})$ とおいてみる。この分布に対して、条件付エントロピー

$$H_i \equiv -\sum_{j \in i \cup \mathcal{N}_i} p(j|i) \ln p(j|i)$$

はただちに計算できて、$\ln k$ となる。したがって、この分布のパープレキシティー e^{H_i} は、ちょうど k となっている。この例は、パープレキシティーが最近傍ノードの個数と同様の役割を演じることを示している [1]。

i を中心ノードとする近傍グラフのコンパクトさの指標は、非類似度の期待値

$$\langle d_i \rangle \equiv \sum_{j \in \mathcal{N}_i} d_{i,j} p(j|i)$$

として自然に表現できるから、上記パープレキシティーの性質と併せると、分布 $p(j|i)$ は、以下の最適化問題の解として定めればいいことがわかる。

$$\min \langle d_i \rangle \ \text{ s.t. } \ \mathrm{e}^{H_i} = c \quad \text{および式 (1)} \tag{2}$$

H_i と規格化条件 (1) に対して未定係数を使い、$p(j|i)$ で微分することで、結局

$$p(j|i) = \frac{1}{Z_i} \mathrm{e}^{-\frac{d_{ij}}{\sigma_i}} \tag{3}$$

を得る。ただし分配関数 Z_i は次のように定義される。

$$Z_i \equiv 1 + \sum_{l \in \mathcal{N}_i} \mathrm{e}^{-\frac{d_{il}}{\sigma_i}} \tag{4}$$

未定係数 σ_i は、もし c が与えられていれば、パープレキシティーの条件から定めることができる。あるいは、これが $d_{i,j}$ の定義に含められていると考えて $\sigma_i = 1$ と置いてしまうのも実用的には便利である。

3.2 異常度の定義

テスト用データと見本データから、分布 $p(j|i)$ と $\bar{p}(j|i)$ が求まったとしよう。近傍保存原理に従えば、下記の量

[1] 確率的近傍グラフの文脈では、この事実は Hinton および Roweis により初めて指摘された [3]。

の差は、システムが正常に動いていれば小さいはずである。

$$e_i(\mathcal{N}_i) \equiv \sum_{j\in\mathcal{N}_i} p(j|i) \quad (5)$$

$$\bar{e}_i(\mathcal{N}_i) \equiv \sum_{j\in\mathcal{N}_i} \bar{p}(j|i) \quad (6)$$

明らかに e_i は、中心ノードとそれを取り巻く近傍ノードとの結合の緊密さを表現する量である。ただし、上式では、k 近傍を、テストデータに基づいて決めている。同様に、見本データの k 近傍集合 $\bar{\mathcal{N}}_i$ を用いて、$e_i(\bar{\mathcal{N}}_i)$ と $\bar{e}_i(\bar{\mathcal{N}}_i)$ も定義することができる。e_i などは確率値であり、その差は確率の変化分という明確な意味を持つ。更に都合の良いことに、

$$0 \le e_i \le \frac{k}{k+1} \quad (7)$$

が成り立つ。同じ式は $\bar{e}_i$ についても成り立つ。下限は中心ノード i が完全に他と無相関な場合であり、上限は、近傍ノードと完全相関である場合を表している。

これらの量を用いて、ノード i（したがってセンサー i）の異常度 E を次のように定義する。

$$E \equiv \max\left\{\left|e_i(\mathcal{N}_i) - \bar{e}_i(\mathcal{N}_i)\right|, \left|e_i(\bar{\mathcal{N}}_i) - \bar{e}_i(\bar{\mathcal{N}}_i)\right|\right\} \quad (8)$$

これがわれわれの**異常度の定義**である。明らかにこれは式 (7) と同じ上下限を持ち、それ自体確率値の変化という意味を持っている。これは値の解釈を大変容易にする。

3.3 異常度の性質

自己結合確率の役割。われわれの確率的近傍グラフの定式化は、値が不安定なセンサーによるノイズを自動的に除去する機能を備えている。そのようなセンサー j は、ある中心ノード i に対し、$d_{i,j} \ll 1$ を満たすため $(j \neq i)$、確率重みは $p(i|i)$ に集中する。結果として、k 最近傍グラフのタイトさは低く、異常度は常に小さくなる。この性質により、不安定な変数がシステムに混じっていても、ロバストに異常度を計算できる。自己結合項は従来の確率的近傍の定式化 [3, 2] では含めないのが普通であるが、ここでは本質的な役割を演じている。

パラメターの選択。σ_i を別にすれば、異常度を計算するために指定すべき唯一のパラメターが、最近傍ノードの数 k である。これは理論的には、タイトなクラスターのサイズとおおよそ等しく選ぶのが正しい。ただ、一般にこの値を最適に決めるのは簡単ではなく、それ自体研究の対象と言える [14]。われわれの応用では、$k = 2$ または 3 と選ぶのが通例である。実応用では、センサーの名称などから直感的に k の値を決められる場合も多い。

計算量。相違度行列と k 最近傍グラフを計算するためには、それぞれ TN^2 と kN^2 の計算量が必要である。もし N が O(10^3) 以上になるようなことがあれば、たとえば近傍探索のための各種の工夫が必要になる。以下の応用では、しかし、素朴にこれらの計算量を甘受しても特に支障はない。

4 実験

この節では性質の異なる二つのデータを使って、われわれの手法の諸性質を調べる。ひとつは図 1 に示したような非常に動的かつ異種混合的なデータで、もうひとつは各センサー信号に類似性が比較的高く、変化も弱い場合のデータである。冒頭に掲げたような条件で変化解析を行える手法はまだ確立していないため、他手法と意味のある比較実験を行うのは難しいが、ここでは定性的に PCA および MDS ベースの手法 [8, 4] との比較を行う。その他の実験結果については別論文 [6] も参照されたい。

4.1 データセット

MotorCurrent データ。UCR アーカイブ [7] から入手できる MotorCurrent というデータから、図 3 のデータを作成した。見本データとしては “healthy”を使い（$N = 20$、$T = 1,500$）、テストデータは、見本データの第 1、第 2 番目の時系列を “1 broken bar”のそれと置き換え、長さ $\bar{T} = 1,000$ の部分をランダムに切り出すことで作成した。便宜上、下から上に、x_1 から x_{20} までの名前を与える。われわれのゴールは、データについてのこのような知識なしに各変数の異常度を計算し、x_1 と x_2 を異常センサーとして同定することである。

Machine データ。ある機械システムに 61 個のセンサーを取り付け、およそ 1 分間のデータを採取し、図 1 のように見本データとテストデータを作成した。測定の時間分解能は、前処理によりすべての時系列について 0.2 秒とした（従って $N = 61$、$T \approx 300$）。テストデータについては、平均と分散がほとんど変わらない 3 つのセンサーの配線を入れ換えることでエラーを挿入した。おおむね、図 1 の上から 2 から 4 番目の変数の内部で巡回置換したものと同様である。このような入れ換えミスは、センサー自体は死んでいるわけではないので見た目にはほとんどわからず、検知が難しい。われわれのゴールは、データについてのこのような知識なしに各変数の異常度を計算し、配線間違いを起こしているセンサーを同定することである。

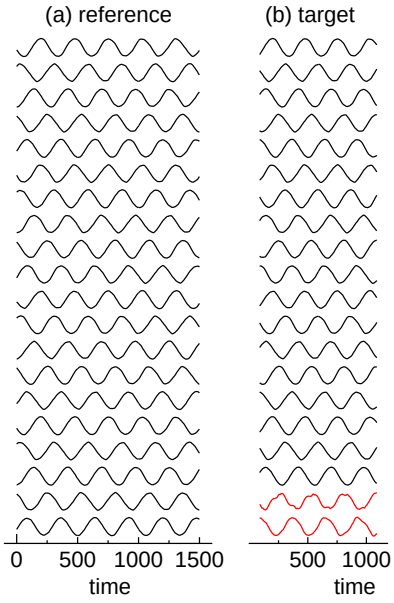

図 3: MotorCurrent データ。

4.2 解析結果

Sammon マップ。データのグローバルな傾向をつかむため、各データについて Sammon マップを計算し、図 4 と図 5 に示した。入力は時系列同士のユークリッド距離行列である。Sammon マップ [1] は MDS の一種であるが、いわゆる古典 MDS よりも視覚的に質の高いマップを作るとされる。古典 MDS が PCA と等価であることから、見本とテストのマップを見比べることは、「PCA による特徴抽出がデータの本質的構造を捉えられるかどうか」を見るという意味を持つ。

MotorCurrent データの Sammon マップ（図 4）では、時系列における位相のずれ具合から予期されるように、5 つの明瞭なクラスターが現れている。見やすいように、異常を仕込んだ x_1 と x_2 をそれぞれ ‘+’ と ‘×’ で区別している 。テストデータ (b) では、異常を挿入した変数はクラスターから若干外れているものの、基本的なクラスターの構造は保持されている。すなわち、このデータのように比較的おとなしく一様なデータでは、データのグローバルな構造は保持されている。

一方、Machine データの Sammon マップ（図 5）では、見本データ (a) とテストデータ (b) に配置の類似性を見出すのは困難である。これも見やすいように、配線置換をしたセンサーを、‘+’, ‘×’, and ‘∗’ で区別している。この例が明らかに示しているように、高度に動的なデータでは、変数間のグローバルな構造が、異なる実験データ間で保持されると想定することはできず、たとえば PCA で抽出した主構造を元に異常を定義するというような方法 [8] には原理的な困難がある。MDS 座標のずれから

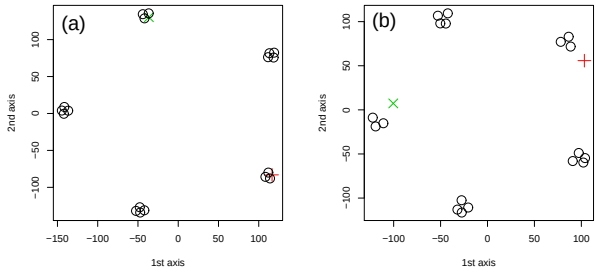

図 4: MotorCurrent データの Sammon マップ。(a) 見本データ、(b) テストデータ。

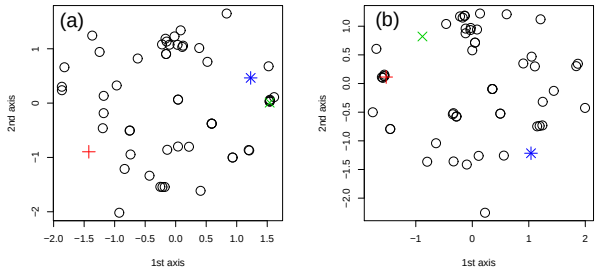

図 5: Machine データの Sammon マップ。(a) 見本データ、(b) テストデータ。

異常度を推定するというような方法も同様である [4]。

異常度。図 6 に MotorCurrent データの異常度を示す。ただし $k = 3$ とした（クラスターサイズの観察から自然な選択である）。また、数値的不安定性を避けるため、$\sigma_i = 1$ と固定した。図 4 から予期されるように、この場合、近傍グラフのメンバーシップには変化がなく、異常度の値自体は大きくない。それでも図から、x_1 と x_2 が明確に高い異常度を示していることがわかる。ただ、これらの変数がクラスターから若干外れた「反動」で、x_6 など、同じクラスターに入っていた変数のスコアにもインパクトが出ている。

図 7 は Machine データの計算結果である。実際の実験では 11 個のエラーパターンに対して、25 個の見本-テスト対を作った。これは 11 個の中で最も悪い（つまり異常センサーと他のコントラストが最も低い）エラーパターンを選び、その中から見本-テスト対ひとつを取って示したものである。実際にエラーを呈していたセンサーを点線で囲った。$k \geq 4$ でスコアは k に対して鈍感であり、明瞭に異常センサーを検知していることがわかる。詳しく調べると、この 4 という数字は、図 5 (b) において ‘+’ の属するクラスターのサイズに関係している。この実験結果は、グローバルな構造が保存されないような

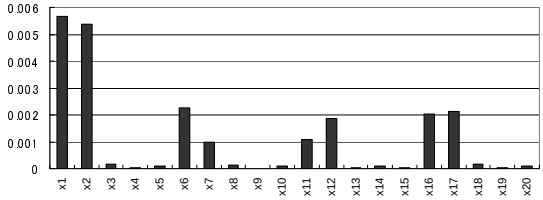

図 6: MotorCurrent データの異常度スコア。

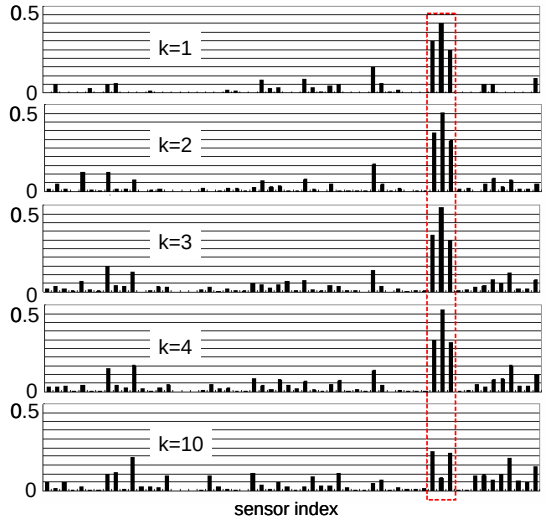

図 7: Machine データの異常度スコアとその k 依存性。

過酷な状況においても、近傍保存原理に基づくわれわれのアプローチがうまく動いていることを示している。

5 まとめ

実世界のセンサーデータは、異種混合的であったり、極めて動的であったりと、非常に「汚い」特徴を持っている。本論文では、近傍保存原理という素朴な原理を、確率的近傍グラフの定式化を流用して実装することにより、そういう汚さに負けず非常にロバストに異常を検知できることを見た。

参考文献

[1] T. F. Cox and M. A. A. Cox. *Multidimensional Scaling, 2nd ed.* Chapman and Hall, 2001.

[2] J. Goldberger, S. Roweis, G. Hinton, and R. Salakhutdinov. Neighbourhood component analysis. In *Advances in Neural Information Processing Systems, 17*, pages 513–520, 2005.

[3] G. Hinton and S. T. Roweis. Stochastic neighbor embedding. In *Advances in Neural Information Processing Systems, 15*, pages 833–840, 2003.

[4] T. Idé and K. Inoue. Knowledge discovery from heterogeneous dynamic systems using change-point correlations. In *Proc. SIAM Intl. Conf. Data Mining*, pages 571–575, 2005.

[5] T. Idé and H. Kashima. Eigenspace-based anomaly detection in computer systems. In *Proc. ACM SIGKDD Intl. Conf. Knowledge Discovery and Data Mining*, pages 440–449, 2004.

[6] T. Idé, S. Papadimitriou, and M. Vlachos. Computing correlation anomaly scores using stochastic nearest neighbors. In *Proc. IEEE Intl. Conf. Data Mining*, page (to appear), 2007.

[7] E. Keogh and T. Folias. The UCR time series data mining archive [`http://www.cs.ucr.edu/~eamonn/TSDMA/index.html`]. 2002.

[8] S. Papadimitriou, J. Sun, and C. Faloutsos. Streaming pattern discovery in multiple time-series. In *Proc. Intl. Conf. Very Large Data Bases*, pages 697–708, 2005.

[9] S. Papadimitriou, J. Sun, and P. Yu. Local correlation tracking in time series. In *Proc. IEEE Intl. Conf. Data Mining*, pages 456–465. IEEE, 2006.

[10] J. Sun, H. Qu, D. Chakrabarti, and C. Faloutsos. Neighborhood formation and anomaly detection in bipartite graphs. In *Proc. IEEE Intl. Conf. Data Mining*, pages 418–425, 2005.

[11] J. Sun, Y. Xie, H. Zhang, and C. Faloutsos. Less is more: Compact matrix decomposition for large sparse graphs. In *Proc. SIAM Intl. Conf. Data Mining*, 2007.

[12] J. Takeuchi and K. Yamanishi. A unifying framework for detecting outliers and change points from time series. *IEEE Trans. Knowledge and Data Engineering*, 18(4):482 – 492, 2006.

[13] K. Yamanishi, J. Takeuchi, G. Williams, and P. Milne. On-line unsupervised outlier detection using finite mixtures with discounting learning algorithms. *Data Mining and Knowledge Discovery*, 8(3):275–300, 2004.

[14] X. Yang, S. Michea, and H. Zha. Conical dimension as an intrinsic dimension estimator and its applications. In *Proc. SIAM Intl. Conf. Data Mining*, 2007.