Vol.2016-MPS-111 No.20
2016/12/12

近傍エゴネットワークにおける多段多数決に基づくクラス分類手法の提案

大久保 好章[1,a)]　原口 誠[1,b)]

概要：本稿では，クラスラベルが未知のクエリオブジェクトのクラスを予測するクラス分類問題のための伝統的な手法である k-近傍法の改良について議論する．具体的には，近傍パラメータ k が予測結果に対して直接的に及ぼす影響を緩和すべく，クエリの近傍オブジェクトのラベルを，その周辺の隣接構造をもとに必要に応じて修正することで，パラメータ設定の違いや近傍の例外的なラベルに影響されない安定した予測結果を得ることが期待できる手法を提案する．

Improving k-NN Method by Modifying Class Labels of Neighbors Based on Their Adjacency Structures

Yoshiaki OKUBO[1,a)]　Makoto HARAGUCHI[1,b)]

1. はじめに

本稿では，**クラス分類問題**のための伝統的手法である k-**近傍法** [1], [2] の改良案について議論する．

クラス分類は，**機械学習**における主要なタスクのひとつであり，クラスラベルが既知のオブジェクト集合をもとに，クエリとして与えられたクラスが未知のオブジェクトが属する適切なクラスを予測する**教師あり学習問題**のひとつとされている．その身近な具体例として，**手書き文字の認識** [3]，**スパムメイルのフィルタリング** (**識別**) [4]，**音楽のジャンル分類** [5] 等が挙げられる．

クラス分類手法として，k-**近傍法** (*k-Nearest Neighbor Method*) が古くから知られている [1], [2]．そこでは，クエリと最も距離の近い k のオブジェクトに注目し，それらの多数が属するクラスを予測結果とする**多数決の原理**が採用されている．基本アイデアは極めて単純かつ素直なものであるが，その妥当性は経験的にも認められ，実際，オブジェクトに関する十分な知識を持たない (あるいは仮定できない) 場合には，事実上この様な単純な手法に頼らざるを得ない．こうした背景もあり，k-近傍法は今もなお有用な手法として様々な応用において利用され，その改良が続けられている (例えば [6], [7], [8] 等)．

一方で，その単純さ故に生じる問題もある．具体的には，唯一の近傍パラメータ k の違いが，しばしば予測結果を大きく左右してしまう．しかし，k の適切な値を事前に知ることは多くの場合困難であり，ユーザはその設定のための試行錯誤を余儀なくされる．k の値を大きくすることで，より安定した多数決を期待できる反面，クエリと離れたオブジェクトがその近傍に含まれる危険性もあり，予測結果に悪影響を与え兼ねない．また，k の値を小さくし過ぎると，極少数の近傍オブジェクトだけを頼りに予測することになり，そこに例外的なオブジェクトが含まれていた場合，予測を誤る可能性が高くなる．

本稿では，近傍パラメータ k が分類の予測結果に及ぼす直接的な影響を緩和するために，クエリの近傍オブジェクトに付与された元のクラスラベルを，その周辺の隣接構造から定まる**仮説ラベル**と比較し，両者が一致しない場合は，元のラベルを仮説ラベルに修正した後，多数決によってクエリが属するクラスを予測する手法を提案する．

[1] 北海道大学大学院情報科学研究科
Graduate School of Information Science and Technology, Hokkaido University
[a)] yoshiaki@ist.hokudai.ac.jp
[b)] mh@ist.hokudai.ac.jp

2. 準備

本稿では，多重辺，および，自己ループ辺を含まない**単純無向グラフ**を扱うものとし，以降ではこれを単にグラフと呼ぶ．

頂点集合 V，**辺集合** $E \subseteq V \times V$ のグラフを $G = (V, E)$ と表記する．任意の頂点 $u, v \in V$ について，$(u, v) \in E$ である時，u と v は G において互いに**隣接**すると言う．頂点 $v \in V$ に隣接する頂点集合を $N_G(v)$ で参照する．

グラフ $G = (V, E)$ の頂点集合 $X \subseteq V$ について，$G[X] = (X, E \cap (X \times X))$ で定義されるグラフを，X による G の**誘導部分グラフ**と呼ぶ．特に，$G[X]$ における任意の異なる頂点のペアが隣接する時，$G[X]$ を G の**クリーク** (**完全部分グラフ**) と呼ぶ．表記を簡略化するため，クリーク $G[X]$ を，それを誘導する頂点集合 X により参照する．

グラフ $G = (V, E)$ の頂点 $v \in V$ について，その隣接頂点集合 $N_G(v)$ による誘導部分グラフ $G[N_G(v)]$ を，G における v の**エゴネットワーク** [9] と呼び，$G_{ego}(v)$ と表記する．

3. k-近傍法によるクラス分類

3.1 クラス分類問題

クラスラベルの集合を $\mathcal{CL} = \{C_1, \ldots, C_n\}$ とする．クラスラベルが未知のオブジェクトに $\mathcal{CL}$ 中のいずれかのラベルを適切に割り当てる問題を n-**クラス分類問題**と呼ぶ．特に，$n > 2$ の場合は**多クラス分類問題**とも呼ばれる．n-クラス分類問題を解く伝統的な手法として k-近傍法が知られており，本稿ではその改良について議論する．

3.2 k-近傍法

すべてのオブジェクトの集合を $\mathcal{U}$ とし，$\mathcal{U}$ 上の**距離関数**を $dist : \mathcal{U} \times \mathcal{U} \mapsto \mathcal{R}$ とする．いま，あるオブジェクト $x \in \mathcal{U}$ と，オブジェクト集合 $\mathcal{D} \subseteq \mathcal{U}$ を考える．(正の) 自然数を定義域とする**近傍パラメータ**を $k \in \mathbb{N}^+$ とした時，x に最も距離が近い $\mathcal{D}$ 中の k のオブジェクトを，$\mathcal{D}$ における x の k-**近傍**と呼び，$kNN^x_{\mathcal{D}}$ で参照する．

クラスラベルが未知のオブジェクト $q \in \mathcal{U}$ を**クエリオブジェクト**，または単に**クエリ**と呼び，オブジェクト集合 $\mathcal{D} \subseteq \mathcal{U}$ をもとに，クエリ q のラベルを予測する問題を考える．ここで，各オブジェクト $x \in \mathcal{D}$ には，$\mathcal{CL}$ 中のいずれかひとつのクラスラベル C_i が付与されているとし，x を C_i-**オブジェクト**と呼び，x のラベルを $label(x)$ で参照する．また，任意のラベル $C_i \in \mathcal{CL}$ について，$\mathcal{D}$ 中の C_i-オブジェクトの集合 $\{x \in \mathcal{D} \mid label(x) = C_i\}$ を $\mathcal{D}(C_i)$ で参照する．

k-**近傍法** (k-*Nearest Neighbor Method*) [1], [2] は，$\mathcal{D}$ に

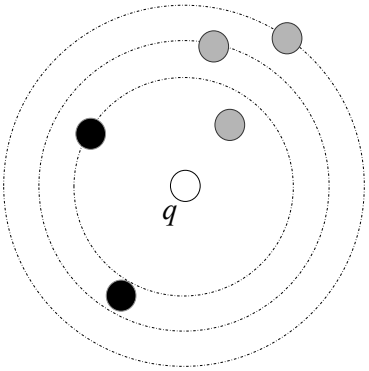

図 1 k-近傍法におけるパラメータ設定の影響

おけるクエリ q の k-近傍 $kNN^q_{\mathcal{D}}$ のラベル情報に基いて q の (適切と思われる) ラベルを予測する手法であり，具体的には，$kNN^q_{\mathcal{D}}$ 中のオブジェクトに付与されたラベルの**多数決**による．すなわち，q のラベルは

$$\arg \max_{C_i \in \mathcal{CL}} \{|kNN^q_{\mathcal{D}}(C_i)|\}$$

であると予測される．$k = 1$ の場合は特に**最近傍法**と呼ばれる．

3.3 k-近傍法の問題点

k-近傍法によるクエリラベルの予測結果は，言うまでもなく近傍パラメータ k に強く依存する．経験的には，k は比較的小さな値に設定されることが多いが，例えば，図 1 においてクエリ q のクラスラベル (色) を予測する場合，$k = 1$ ではグレー，$k = 3$ では黒，$k = 5$ ではグレーとなり，その設定が予測結果に及ぼす影響は極めて大きい．また，仮に適当な k を設定できたとしても，k-近傍の (一部の) オブジェクトに付与されたラベルが**例外的**なものであった場合，それが結果に大きく影響する可能性があることは容易に想像できる．

4. 近傍エゴネットワークにおける多段多数決による k-近傍法の改良

本節では，k-近傍法におけるパラメータ設定に起因する予測結果への直接的な影響を緩和する手法を提案する．具体的には，クエリの k-近傍オブジェクトそれぞれについて，その周辺を考慮したラベル修正を (必要に応じて) 行い，その修正ラベルをもとにクエリのラベルを予測する k-近傍法の改良について議論する．

4.1 基本アイデア

いま，$\mathcal{CL} = \{C_1, \ldots, C_n\}$ 中のいずれかのクラスラベルが付与されたオブジェクト集合 $\mathcal{D}$ をもとに，クエリオブジェクト q のラベルを予測する．

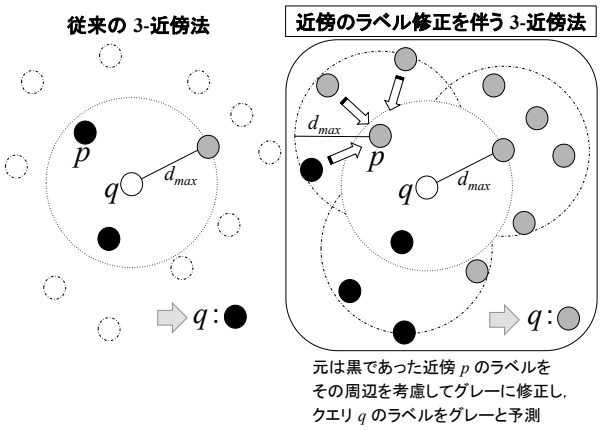

図 2 ラベル修正を伴う k-近傍法の改良

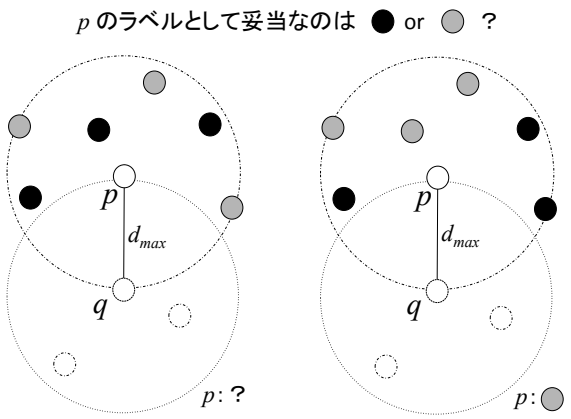

図 3 p の距離-d_{max}-近傍の成す隣接構造の違い

近傍パラメータを $kb \in \mathbb{N}^+$ とし，$\mathcal{D}$ 中の q の k-近傍 $kNN^q_{\mathcal{D}}$ の中で，q から最も離れたオブジェクトまでの距離を d_{max} とする．k-近傍法は，$kNN^q_{\mathcal{D}}$ 中のオブジェクトに付与されたラベルをもとに q のラベルを予測する手法であるが，このことは，q から距離 d_{max} 以内にあるオブジェクトのラベルをもとに q のラベルを予測することに他ならない．ここで，この考え方を q の k-近傍オブジェクト $p \in kNN^q_{\mathcal{D}}$ に対して逆に当てはめると，p のラベル $label(p)$ は，p から距離 d_{max} 以内にあるオブジェクトのラベルと整合がとれるべきと考えることが妥当であろう．すると，$label(p)$ が距離 d_{max} 以内にあるオブジェクトのラベルの多くと異なっている場合，$label(p)$ は例外的なものとも解釈でき，q のラベル予測時に，$label(p)$ をそのまま多数決に用いることは望ましくないと考えられるだろう．こうした場合，本稿では，$label(p)$ を距離 d_{max} 以内のオブジェクトと整合のとれるラベルへと修正し，それを q のラベル予測の多数決に用いるものとする (図 2 参照)．つまり，クエリ q の k-**近傍の周辺**を考慮したラベルの修正を (必要に応じて) 行うことで，k-近傍に付与された例外的なラベルの影響を抑制した多数決が可能となり，結果として，ラベルの予測結果に及ぼす近傍パラメータ k の直接的な影響の緩和が期待できる．

4.2 近傍エゴネットワークにおける多段多数決による近傍ラベルの修正

$\mathcal{D}$ 中の q の k-近傍 $kNN^q_{\mathcal{D}}$ の中で，q から最も離れたオブジェクトまでの距離を d_{max}，つまり，$d_{max} = \max_{x \in kNN^q_{\mathcal{D}}}\{dist(q,x)\}$ とする．いま，q の k-近傍オブジェクト $p \in kNN^q_{\mathcal{D}}$ について，p から距離 d_{max} 以内にある $\mathcal{D}$ 中のオブジェクト集合を p の**距離**-d_{max}-**近傍**と呼び，$W^{d_{max}}_{\mathcal{D}}(p)$ で参照する．すなわち，

$$W^{d_{max}}_{\mathcal{D}}(p) = \{x \in \mathcal{D} \mid dist(p,x) \leq d_{max}\}$$

である．

先に述べた通り，k-近傍法のアイデアに従うと，q の k-近傍オブジェクト p のラベル $label(p)$ は，p の距離-d_{max}-近傍オブジェクトのラベルと整合することが望ましい．その整合性を判断する素朴な方法として，$W^{d_{max}}_{\mathcal{D}}(p)$ 中のオブジェクトのラベルによる多数決結果と $label(p)$ との比較が考えられるが，単純な多数決では $W^{d_{max}}_{\mathcal{D}}(p)$ 中のオブジェクトの分布が考慮されないという問題がある．例えば，p の距離-d_{max}-近傍オブジェクトが図 3 の様に異なって分布する場合，単純な多数決によると両者とも 3 対 3 となり p のラベルを判断しかねるが，それらの分布が示す隣接構造を考慮すると，少なくとも右図の場合はグレーが集中しており，グレーであると考えることが妥当に思える．

本稿ではこうした隣接構造を考慮した多数決を，**距離**-d_{max}-**近傍における多段多数決**により実現し，その結果を元の $label(p)$ と比較することで，一致しない場合は $label(p)$ の修正を試みる．

4.2.1 近傍エゴネットワークの構築

クエリオブジェクト q の k-近傍オブジェクト $p \in kNN^q_{\mathcal{D}}$ について，p の距離-d_{max}-近傍 $W^{d_{max}}_{\mathcal{D}}(p)$ を考える．ここで，$V_p = W^{d_{max}}_{\mathcal{D}}(p)$ を頂点集合，上限距離 d_{max} による V_p 上の隣接関係 $E_p = \{(x,y) \mid x,y \in V_p \text{ and } dist(x,y) \leq d_{max}\}$ を辺集合とする (無向) グラフを $G = (V_p, E_p)$ とすると，G は p の距離-d_{max}-近傍の隣接構造を与える．特に，G における (クエリ q の k-近傍である)p のエゴネットワーク $G_{ego}(p) = G[N_G(p)]$ は，p の周辺に位置するオブジェクトが成す隣接構造を与えるものであり，これを q に関する p-**近傍エゴネットワーク**と呼び，この隣接構造をもとに p のラベル $label(p)$ を必要に応じて修正する．

4.2.2 近傍エゴネットワークの極大クリークに基づく局所ラベルの同定

クエリ q の k-近傍 p を取り巻く p-近傍エゴネットワーク $G_{ego}(p)$ におけるクリークを考える．$G_{ego}(p)$ のクリーク Q を構成する頂点 (オブジェクト) は互いに隣接関係にあることから，これらのラベルによる多数決の結果を Q の

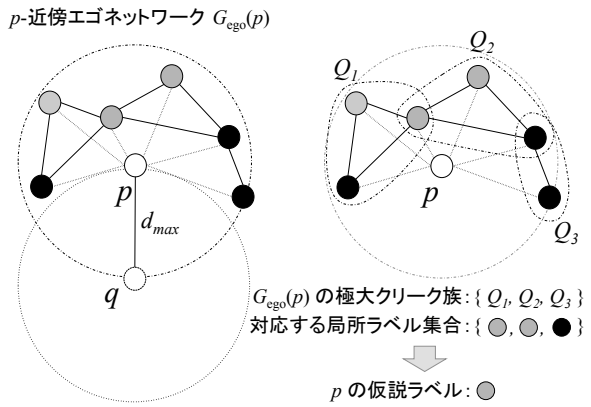

図 4 多段多数決による近傍 p の仮説ラベル同定

ラベルとすることには合理性があろう．これにより，p 周辺の**局所的な隣接構造**から定まる p の局所ラベルを得ることができる．一般に，$G_{ego}(p)$ には複数のクリークが存在するが，クリークの部分グラフはまたクリークであるから，ここでは包含関係のもとで極大なもの，すなわち，**極大クリーク** [10], [11] のみを考え，各極大クリークについて p の局所ラベルをひとつ同定するものとする．

形式的には，p-近傍エゴネットワーク $G_{ego}(p)$ の極大クリーク族を $\mathcal{MCQ}(G_{ego}(p)) = \{Q_1, \ldots, Q_\ell\}$ とする．任意の極大クリーク $Q_i \in \mathcal{MCQ}(G_{ego}(p))$ について，その構成頂点のラベルによる多数決結果を Q_i に基づく p の**局所ラベル**と呼び，$l\text{-}label(Q_i)$ で参照する．つまり，

$$l\text{-}label(Q_i) = \arg\max_{C \in \mathcal{CL}} \{|Q_i(C)|\}$$

となる．

4.2.3 局所ラベルの多数決による近傍仮説ラベルの同定

クエリ q の k-近傍オブジェクト p について，p-近傍エゴネットワークの極大クリーク族 $\mathcal{MCQ}(G_{ego}(p)) = \{Q_1, \ldots, Q_\ell\}$ から得られる局所ラベル集合 $\{l\text{-}label(Q_1), \ldots, l\text{-}label(Q_\ell)\}$ を考える．各局所ラベル $l\text{-}label(Q_i)$ は，p の周辺の局所的な隣接構造に基づいて得られたラベルであり，ここではそれらの多数決結果を，周辺隣接構造から定まる p の**仮説ラベル**と呼び，$h\text{-}label(p)$ で参照する．すなわち，多重集合 X において出現回数が最も多い要素を $majority(X)$ で参照すると，$h\text{-}label(p)$ は

$$h\text{-}label(p) = majority\left(\{l\text{-}label(Q_i)\}_{i=1}^{\ell}\right)$$

で与えられる．

p-近傍エゴネットワークの各極大クリークを構成するオブジェクトラベルの多数決により局所ラベルが決まり，それら局所ラベルの多数決によって p の仮説ラベルが決まる．この様に，p の仮説ラベル $h\text{-}label(p)$ は，p-近傍エゴネットワークにおける**多段多数決**により決定される (図 4 を参照)．

4.2.4 近傍ラベルの修正

クエリ q の k-近傍オブジェクトを $p \in kNN_{\mathcal{D}}^{q}$ とし，p-近傍エゴネットワークにおける多段多数決により，その仮説ラベル $h\text{-}label(p)$ が得られているとする．仮説ラベルは p 周辺の局所的な隣接構造を考慮したラベルであり，p に (予め) 付与された $label(p)$ と一致していることが望ましいであろう．逆に，もし $label(p) \neq h\text{-}label(p)$ の場合，ここでは，$label(p)$ は例外的に付与されたものと考え，仮説ラベル $h\text{-}label(p)$ へと修正するものとする．

4.2.5 近傍ラベルの修正を伴うクエリのラベル予測

従来の k-近傍法では，クエリ q の k-近傍オブジェクト $p \in kNN_{\mathcal{D}}^{q}$ のラベル $label(p)$ を用いて q のラベルを予測する．ここでは，仮説ラベル $h\text{-}label(p)$ との比較により，必要に応じて $label(p)$ を修正した上で，q のラベルを予測することを考える．

形式的には，任意の k-近傍オブジェクト $p \in kNN_{\mathcal{D}}^{q}$ について，多数決に実際に用いるラベルを与える関数 $Label : \mathcal{D} \mapsto \mathcal{CL}$ を次の通り定義すればよい．

$$Label(p) = \begin{cases} label(p), & \text{if } label(p) = h\text{-}label(p), \\ h\text{-}label(p), & \text{otherwise.} \end{cases}$$

これらラベルを用いてクエリ q のラベル $ClassLabel(q)$ を

$$ClassLabel(q) = majority\left(\{Label(p)\}_{p \in kNN_{\mathcal{D}}^{q}}\right).$$

と予測する．

4.3 ラベル修正を伴う拡張 k-近傍法アルゴリズム

これまでの議論に基づくラベル修正を伴う拡張 k-近傍法アルゴリズムを図 5 に示す．

図中，MaximalCliques は，所与のグラフにおける極大クリークを列挙する手続きであり，その代表的なものとして CLIQUES [10] が知られている．

5. おわりに

本稿では，k-近傍法において，分類予測結果に対して近傍パラメータが直接的に及ぼす影響を緩和すべく，クエリの近傍オブジェクトのラベル修正を伴う改良法について議論した．近傍エゴネットワークにおける多段多数決により，近傍オブジェクトの元のラベルが (必要に応じて) 仮説ラベルに修正され，それら修正後のラベルによる多数決をとることで，パラメータ設定の違いや近傍の例外的なラベルに影響されない安定した予測結果を得ることが期待できる．

現在，提案アルゴリズムの実装作業を進めている段階であり，近傍パラメータの変化に伴う分類精度の挙動を観察することで，従来手法の問題点が緩和されていることを実験的に確認したい．また，Support Vector Machine をはじめとする他のクラス分類手法との比較を通して，提案手

procedure KMensWithLabelModification($\mathcal{D}$, k, q):
  [**Input**] $\mathcal{D}$: a set of objects with class labels in $\mathcal{CL} = \{C_1, \ldots, C_n\}$.
    k: an integer for k-nearest neighbors.
    q: a query object without class label.
  [**Predict**]: a class label of q.
  begin
    if there exists an object $x \in \mathcal{D}$ such that $dist(q, x) = 0$ **then**
      return $label(x)$;
    endif
    $kNN \leftarrow k$-nearest neighbors of q in $\mathcal{D}$;
    $d_{max} = \max_{x \in kNN}\{dist(q, x)\}$;
    $\mathcal{L} = \emptyset$; // as a multi-set
    for each $p \in kNN$
      create p-ego network $G_{ego} = (V, E)$, where
        $V = \{x \in \mathcal{D} \mid dist(p, x) \leq d_{max}\}$ and
        $E = \{(x, y) \mid x, y \in V$ and $dist(x, y) \leq d_{max}\}$.
      $\mathcal{Q} \leftarrow$ MaximalCliques(G_{ego});
      $L\text{-}label = \emptyset$; // as a multi-set
      for each $Q \in \mathcal{Q}$
        $labelQ = majority(\{label(x)\}_{x \in Q})$; // as $l\text{-}label(Q)$
        $L\text{-}label \leftarrow L\text{-}label \cup \{labelQ\}$;
      endfor
      $H\text{-}label = majority(L\text{-}label)$; // as $h\text{-}label$(p)
      if $label(p) = H\text{-}label$ **then** // modifying original label if necessary
        $\mathcal{L} \leftarrow \mathcal{L} \cup \{label(p)\}$;
      else
        $\mathcal{L} \leftarrow \mathcal{L} \cup \{H\text{-}label\}$;
      endif
    endfor
    return $majority(\mathcal{L})$;
  end

図 5 ラベル修正を伴う拡張 k-近傍法アルゴリズム

法の有効性，および，問題点も検証したい．それら結果については改めて報告する．

参考文献

[1] Cover, T. M. and Hart, P. E.: Nearest Neighbor Pattern Classification, IEEE Transactions on Information Theory, 13(1), pp. 21 – 27, 1967.
[2] Fix, E. and Hodges Jr., J. L.: Discriminatory Analysis - Nonparametric Discrimination: Consistency Properties, International Statistical Review/Revue Internationale de Statistique, 57(3), pp. 238 – 247, ISI, 1989.
[3] Smith, S. J., Bourgoin, M. O., Sims, K. and Voorhees, H. L.: Handwritten Character Classification Using Nearest Neighbor in Large Databases, IEEE Transactions on Pattern Analysis and Machine Intelligence, 16(9), pp. 915 – 919, 1994.
[4] Jiang, S., Pang, G., Wu, M. and Kuang, L.: An Improved k-Nearest-Neighbor Algorithm for Text Categorization, Expert Systems with Applications, 39(1), pp. 1503 – 1509, 2012.
[5] Pampalk, E., Flexer, A., Widmer, G. Improvements of Audio-Based Music Similarity and Genre Classificaton, Proc. of the 6th International Conference on Music Information Retrieval - ISMIR'05, pp. 628 – 633, 2005.
[6] Samanthula, B. K., Elmehdwi, Y. and Jiang, W.: k-Nearest Neighbor Classification over Semantically Secure Encrypted Relational Data, IEEE Transactions on Knowledge and Data Engineering 27(5), pp. 1261 – 1273, 2015.
[7] Zhou, Z., Wen, C. and Yang, C.: Fault Detection Using Random Projections and k-Nearest Neighbor Rule for Semiconductor Manufacturing Processes, IEEE Transactions on Semiconductor Manufacturing, 28(1), pp. 70 – 79, 2015.
[8] Nowak, B. A., Nowicki, R. K., Woźniak, M. and Napoli, C.: Multi-Class Nearest Neighbour Classifier for Incomplete Data Handling, Proc. of the 14th International Conference on Artificial Intelligence and Soft Computing - ICAISC'15, LNCS-9119, pp. 469 – 480, 2015.
[9] (Wouter de) Nooy, W., Mrvar, A. and Batagelj, V.: Exploratory Social Network Analysis with Pajek (2nd Ed.), Structural Analysis in the Social Sciences 34, Cambridge University Press, 2011.
[10] Tomita, E., Tanaka, A. and Takahashi, H.: The Worst-

Case Time Complexity for Generating All Maximal Cliques and Computational Experiments, Theoretical Computer Science, 363(1), pp. 28 – 42, Elsevier, 2006.

[11] Eppstein, D. and Strash, D.: Listing All Maximal Cliques in Large Sparse Real-World Graphs, Proc. of the 10th Int'l Symposium on Experimental Algorithms - SEA'11, LNCS-6630, pp. 364 – 375, 2011.