Vol.2013-MPS-96 No.21
Vol.2013-BIO-36 No.21
2013/12/12

拡張されたナイーブベイズを用いたメタゲノム配列の系統分類

小松 祐城[1]　石田 貴士[1]　秋山 泰[1,2]

概要：メタゲノム解析では環境中の微生物の分布の解明のため，混在した多種の微生物の DNA 断片配列の生物種毎への分類が行われる．近年，DNA シーケンサのスループットが急速に向上しており，より高速に DNA 断片配列を分類する手法が求められている．そこで，我々は計算量が少なく一定の予測精度が得られるナイーブベイズに注目し，ナイーブベイズを拡張したモデルである naive Bayes with a hidden class (NBH) または hidden naive Bayes (HNB) を用いることにより高速かつ高精度なメタゲノム配列の系統分類手法の開発を目指した．また，HNB に関して分類をより高速にするためにアルゴリズムの改良を試み，UCI repository のデータセットを用いてメタゲノム分類以外のケースについても，その予測精度と計算速度の確認を行った．

キーワード：メタゲノム解析，系統分類，ナイーブベイズ

Metagenomic phylogenetic classification by using improved naive Bayes

Komatsu Yuki[1]　Ishida Takasi[1]　Akiyama Yutaka[1,2]

Abstract: In metagenomic analysis, phylogenetic classification of DNA fragments from a variety of microorganisms is often performed to understand the taxonomic composition of microbial communities. In recent years, with the improvement of DNA sequencer's throughput, a faster classifier for metagenomic reads is required. In this research, we focused on naive Bayes which can classify in short time with sufficient accuracy, and aimed to develop a faster and accurate classification method by using improved naive Bayes, naive Bayes with a hidden class (NBH) and hidden naive Bayes (HNB). In addition, we tried to improve algorithm of HNB for a faster classification, and checked the performance and computing time by using UCI reoisitory dataset.

Keywords: metagenomic analysis, phylogenetic classification, naive Bayes

1. はじめに

近年，DNA シーケンサのスループットの向上により短時間で大量のゲノム情報が得られるようになった結果，メタゲノム解析という解析が盛んに行われている．メタゲノム解析とは，土壌や腸内等の環境中に生息する多種の微生物の DNA を分離，培養を経ずに直接 DNA を読み取り，環境中に生息する微生物や遺伝子の分布を解析する研究である．現在，環境中に生息する微生物の 99% は未知であると言われており，メタゲノム解析により微生物群集の生態の理解や微生物が持つ膨大な未知の遺伝子資源の発見が期待されている．このようにメタゲノム解析において微生物種の分布の解明は目的の一つであるが，メタゲノム解析では環境中に生息する多数の微生物から DNA をまとめて抽出するため，様々な微生物種の DNA 断片配列が混在したデータが得られる．そのため，メタゲノム解析ではこれらの DNA 断片配列を配列情報をもとに，それらがどの微

[1] 東京工業大学 大学院情報理工学研究科 計算工学専攻
Graduated School of Information Science and Engineering, Tokyo Institute of Technology
[2] 東京工業大学 情報生命博士教育院
Education Academy of Computational Life Sciences, Tokyo Institute of Technology

生物種に由来するものであるか系統的な分類が行われる．具体的には，生物種は系統的に分類されており，phylum，class，order，family，genus といったような階層性を持っていることを利用し，各階層の分岐群への分類が行われている．

DNA 断片配列の配列情報をもとに系統的な分類を行う様々な手法が提案されているが，それらは配列相同性に基づく分類手法と配列組成に基づく分類手法に大別される．

配列相同性に基づく分類手法では，塩基配列の類似度 (配列相同性) を直接データベース中の配列と比較することにより分類を行う．MEGAN [1] というソフトウェアでは塩基配列の類似度の比較において一般に用いられるソフトウェアである BLAST [2] を利用し，メタゲノム配列の系統分類を行っている．この配列相同性に基づく分類手法は，一般に分類の精度が高いが，DNA の断片配列毎に膨大なデータベースと比較する必要があり，計算量が大きくなるという問題がある．

配列組成に基づく分類手法では，DNA 断片配列から特徴を抽出し，その特徴の類似度に基づき分類を行う．現在，配列組成に基づく分類手法の多くが抽出した特徴ベクトルをもとに機械学習を用いて分類を行っている．一般に機械学習を用いた分類手法は配列相同性に基づく分類手法と比べて分類の精度は低いが，一度モデルを構築すれば分類の際の計算量はそれほど大きくならない傾向がある．なお，機械学習の分類器に関する多くの研究ではモデルを構築するための学習時間の短縮を課題としているが，メタゲノム配列の系統分類においては分類するデータ数が非常に多いため，学習時間の短縮よりも予測時間の短縮が大きな課題である．現在，短時間で大量の DNA 配列が読み取られているが，今後も更なる DNA シーケンサのスループットの向上により分類が必要な DNA 断片配列が増加し続けることが考えられるため，高速な分類手法が求められている．そこで，本研究では高速な分類が期待できる機械学習を用いた分類手法に焦点を当てた．

機械学習を用いたメタゲノム配列の系統分類の手法として，代表的なものに support vector machine (SVM) を用いて系統分類を行う Phylopythia [3] やナイーブベイズを用いて系統分類を行う手法 [4] が提案されている．SVM とナイーブベイズは分類手法として様々な分野で利用されている．SVM は一般に予測精度は高いが，基本的には 2 クラス分類器であるためアルゴリズムによってはクラス数が多い問題を扱う場合や問題が複雑であり線形カーネルが使えない場合に分類があまり高速ではないとされている．一方，ナイーブベイズは高速な分類が可能であるが予測精度があまり高くないとされている．

そこで，我々は高速に分類することができるナイーブベイズに注目し，拡張されたナイーブベイズをメタゲノム配列の系統分類へ応用することで，高速かつ高精度な分類手法の開発を目指した．そのため，本研究では他の分野のいくつかのデータセットに対し，ナイーブベイズに比べて予測精度が高いことが報告されている拡張されたナイーブベイズである naive Bayes with a hidden class (NBH) モデル [5] と hidden naive Bayes (HNB) [6] の予測時間を UCI repository の letter データ [7] により検証した．さらに，HNB に関しては分類をより高速にするためにアルゴリズムの改良を試み，letter データを用いて予測精度と予測時間の確認も行った．そして，我々は NBH と HNB を用いたメタゲノム配列の系統分類手法を提案し，SVM を用いた系統分類，ナイーブベイズを用いた系統分類と予測精度と予測時間を比較した．

2. 関連研究

機械学習における最も一般的な分類器の一つにベイジアンネットワークがある．ベイジアンネットワークは属性を表現するノードとその属性の依存関係を表現するアークから構成される．ベイジアンネットワークの最も単純な分類器として，クラス変数を条件とした属性間の条件付き独立性を仮定したナイーブベイズがある．**図 1** にナイーブベイズの構造を示す．ナイーブベイズでは式 (1) のように同時分布を定義する．

$$P(A_1,\ldots,A_n,C) = P(C)\prod_{i=1}^{n} P(A_i|C) \tag{1}$$

ここで，C はクラス，A_i $(i=1,\ldots,n)$ は属性を表す．また，未知事例 $E=(a_1,a_1,\ldots,a_n)$ が与えられたとき，ナイーブベイズによる分類器は式 (2) で定義される．

$$c(E) = \underset{c\in C}{\operatorname{argmax}}\, P(c)\prod_{i=1}^{n} P(a_i|c) \tag{2}$$

ナイーブベイズは計算量が小さく，一定の予測精度を示すが，ナイーブベイズの条件付き独立性の仮定のため属性間に複雑な依存関係があるデータセットにおいてはあまり良い精度が得られない．また，ナイーブベイズは大きなデータセットに対してそれほど良い精度が得られないことが示されている [8]．

ナイーブベイズの条件付き独立性の仮定を緩和するために，属性間の依存関係を明確に表す有向アークを使ってナイーブベイズを構造的に拡張する研究が行われている．以下の節では，本研究でメタゲノム配列の系統分類へ導入する構造的に拡張されたナイーブベイズについて説明する．

3. 手法

3.1 Naive Bayes with a hidden class モデル

Naive Bayes with a hidden class (NBH) モデル [5] は，ナイーブベイズの条件付き独立性の仮定を緩和するために隠れクラス変数 HC を導入したモデルである．**図 2** に NBH の構造を示す．図 2 に示すように，すべての属性は

Vol.2013-MPS-96 No.21
Vol.2013-BIO-36 No.21
2013/12/12

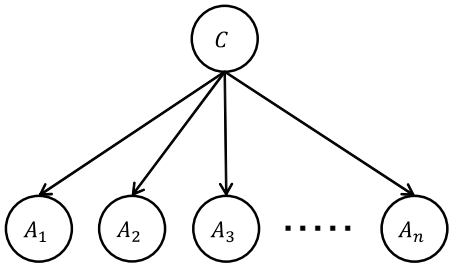

図 1　ナイーブベイズの構造
Fig. 1　The structure of naive Bayes

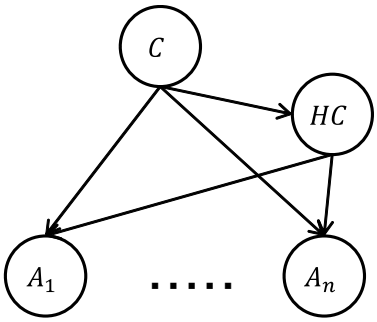

図 2　NBH の構造
Fig. 2　The structure of NBH

クラスとクラスに条件付けされた隠れクラス変数により条件付けされる．

NBH では式 (3) のように同時分布を定義する．

$$P\ (A_1,\ldots,A_n,HC,C|\boldsymbol{\theta}) = \prod_{i=1}^{n} P(A_i|HC,C,\boldsymbol{\theta})P(HC|C,\boldsymbol{\theta})P(C|\boldsymbol{\theta}) \quad (3)$$

ここで，$\boldsymbol{\theta}$ はモデルパラメータ，HC は隠れクラス変数である．

また，未知事例 $E=(a_1,\ldots,a_n)$ が与えられたとき，NBH による分類器は式 (4) のように定義される．

$$\begin{aligned} c(E) &= \operatorname*{argmax}_{c\in C} P(c|a_1,\ldots,a_n,\boldsymbol{\theta}) \\ &= \operatorname*{argmax}_{c\in C} \sum_{hc} P(c|hc,a_1,\ldots,a_n,\boldsymbol{\theta}) \\ &\qquad \times P(hc|a_1,\ldots,a_n,\boldsymbol{\theta}) \end{aligned} \quad (4)$$

この際問題となるのは隠れクラスの数であるが，今回は 10-fold cross-validation で得られる accuracy の平均が最も高い隠れクラスの数を採用することとした．また，NBH においては，モデルパラメータの推定が必要である．本研究では，ベイズ推定の近似計算法である MAP 推定法と変分ベイズ法に基づく変分ベイズ Viterbi Training (VB-VT) [9] によりモデルパラメータの推定を行った．

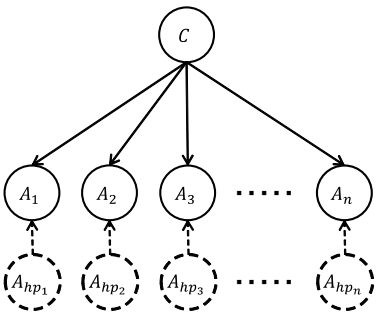

図 3　HNB の構造
Fig. 3　The structure of HNB

3.2 Hidden Naive Bayes

Hidden Naive Bayes (HNB) [6] は，ナイーブベイズの条件付き独立性の仮定を緩和するために，ナイーブベイズの各属性に hidden parent と呼ばれる属性を追加したモデルである．具体的には，hidden parent はある属性のそれ以外のすべての属性からの影響を結合するはたらきがある．

図 3 に HNB の構造を示す．図 3 において，C はクラスであり，すべての属性の親である．また，各属性 A_i は破線の円で示される hidden parent A_{hp_i} $(i=1,2,\ldots,n)$ を持っている．

HNB による同時分布は式 (5) のように定義される．

$$P(A_1,\ldots,A_n,C) = P(C)\prod_{i=1}^{n} P(A_i|A_{hp_i},C) \quad (5)$$

$$P(A_i|A_{hp_i},C) = \sum_{j=1,j\neq i}^{n} W_{ij}\times P(A_i|A_j,C) \quad (6)$$

また，未知事例 $E=(a_1,\ldots,a_n)$ が与えられたとき，HNB による分類器は式 (7) のように定義される．

$$c(E) = \operatorname*{argmax}_{c\in C} P(c)\prod_{i=1}^{n} P(a_i|a_{hp_i},c) \quad (7)$$

$$P(a_i|a_{hp_i},c) = \sum_{j=1,j\neq i}^{n} W_{ij}\times P(a_i|a_j,c) \quad (8)$$

重み W_{ij} $(i,j=1,2,\ldots,n$，ただし，$i\neq j)$ を決定する手法としては，式 (9) で示すように $P(A_i|A_j,C)$ の重みとして，属性 A_i と A_j の間の条件付き相互情報量を使用する．

$$W_{ij} = \frac{I_p(A_i;A_j|C)}{\sum_{j=1,j\neq i} I_p(A_i;A_j|C)} \quad (9)$$

ここで，条件付き相互情報量 $I_P(A_i;A_j|C)$ は式 (10) で定義される．

$$I_p(A_i;A_j|C) = \sum_{a_i,a_j,c} P(a_i,a_j,c)\frac{P(a_i,a_j|c)}{P(a_i|c)P(a_j|c)} \quad (10)$$

3.2.1 HNB の高速化に向けた改良

本研究では高速かつ高精度な分類手法の開発を目指して

いるが，式 (7) と式 (8) を用いた HNB の分類では，ある属性の条件付確率を計算する際に他のすべての属性について考慮する必要があり，属性の数が多くなると考慮する属性のペアが増加し，演算数も増加する．したがって，扱うデータセットの属性の数が多い場合に予測時間が長くなってしまうという問題が考えられる．そこで，我々は考慮する属性のペアの数を減少させることにより予測時間の短縮を試みた．

HNB では条件付き相互情報量により属性間の依存関係を評価しており，この値が高い場合属性間に強い依存関係があるという考え方に基づいている．そこで，条件付き相互情報量の値が小さく，属性間に依存関係が少ないと思われる属性のペアについては式 (8)，(9) において考慮しないことにより演算回数を減少させた．具体的には，条件付き相互情報量に対するしきい値を導入し，しきい値以上の条件付き相互情報量が得られる属性のペアのみ式 (8)，(9) の計算に含めることとした．また，しきい値を大きくすることにより演算回数が大幅に減少する可能性があるが，本研究においては従来の HNB の予測精度をできるだけ低下させないようなしきい値を実験的に選択した．

4. 実験

これまでの研究により他の分野のいくつかのデータセットにおいて，NBH と HNB はナイーブベイズ (naive Bayes, NB) の予測精度を上回る場合があることが示されている [5], [6]．そこで，本研究では UCI repository のデータセットを用いて NBH と HNB の予測精度と予測時間を検証した上で，NBH と HNB をメタゲノム配列の系統分類へ応用し，既存手法として知られるナイーブベイズを用いた系統分類，SVM を用いた系統分類と予測精度と予測時間の比較を行った．また，NBH においては，MAP 推定法によりモデルパラメータを推定するモデル NBH (MAP) と変分ベイズ法に基づく変分ベイズ Viterbi Training によりモデルパラメータを推定するモデル NBH (VB-VT) を比較に用いた．実験に使用した計算機の性能は，Intel Xeon processor X5670 (2.93 GHz)， メモリは 54GB である．

本研究では系統分類に用いる SVM のライブラリとして LIBSVM [10] を利用し，カーネル関数には RBF カーネルを選択した．RBF カーネルを用いた SVM は，正則化パラメータ C とカーネルパラメータ γ の決定が必要である．本実験では，パラメータ調整用のデータセットに対して 10-fold cross-validation を行い，得られる accuracy の平均が最も高いパラメータの組み合わせを SVM の学習パラメータとして利用した．

4.1 UCI repository のデータセットによる予測精度と予測時間の検証

本研究でメタゲノム配列の系統分類へ応用する NBH と HNB は，先行研究において予測精度に関しては他の分類器との比較が行われているが，予測時間についての比較はあまり行われていない．しかし，メタゲノム配列の系統分類においては，大量のデータを高速かつ高精度に分類することが求められているため，NBH (MAP，VB-VT) と HNB の予測精度と予測時間をナイーブベイズ (NB)，SVM と比較した．また，本研究で行った HNB の高速化に向けた改良による効果の検証も行った．

表 1　Accuracy と予測時間の平均
Table 1　Average accuracy and average classification time

分類器	Accuracy の平均 (%)	予測時間の平均 (sec.)
SVM	98.0	1.528
NB	73.6	0.023
HNB	90.1	0.081
HNB (speed-up)	90.1	0.053
NBH (MAP)	90.2	0.456
NBH (VB-VT)	87.4	733.485

予測精度と予測時間の比較には UCI repository 中の letter データ [7] を使用した．letter データはクラス数が 26，属性数が 16，データ数が 20,000 であり，クラス毎のサンプル数は大体等しいものとなっている．また，予測精度として，正しく分類された割合を表す accuracy を用いる．

10-fold cross-validation を行うことにより計測した各分類器の予測精度と予測時間の平均を**表 1** に示す．表 1 に示すように，NBH (MAP，VB-VT) と HNB の accuracy はナイーブベイズの accuracy を上回り，先行研究の結果と一致している．ナイーブベイズと NBH (MAP)，HNB に関しては，SVM と比較すると accuracy は劣るが，高速な分類が可能であることがわかる．また，letter データにおいて，高速化に向けて改良した HNB (speed-up) は，従来の HNB の accuracy を低下させることなく，1.5 倍程度の高速化を達成している．

4.2 実験データ

DNA 断片配列のデータとして，DNA シーケンサのシミュレータである MetaSim [11] を用いて，NCBI データベースにおいて利用できる Bacteria と Archaea の 390 種の生物のゲノム配列から長さ 100 塩基の DNA 断片配列を 319,972 本準備した．また，本実験では genus 層で分類を行い，このデータセットの genus 層の分岐群は全部で 48 群あるため，クラスの数は 48 である．

そして，本研究においては各分類器の入力として，長さ 100 塩基の DNA 断片配列から 3-mer (3 塩基の部分配列) の出現頻度を抽出し属性として用いた．

本研究で用いる分類器において，SVM と NBH (MAP, VB-VT) はパラメータの調整が必要である．本研究では，上記で述べたデータ数が 319,972 であるデータセットの

1/10 の大きさのサブセットをパラメータの調整用のデータセットとして用いた．

4.3 実験の結果

メタゲノム配列の系統分類へ新たに応用する NBH (MAP, VB-VT)，HNB と既に応用されているナイーブベイズ (NB)，SVM の予測精度と予測時間の比較を行った．

本実験で用いるメタゲノム配列のデータセットにおいては，クラス毎のサンプル数に偏りが大きく，サンプル数の最小値と最大値には約 35 倍の差がある．そこで，各クラスのデータ数が不均衡であるデータセットに大して有効である receiver-operator characteristics (ROC) 曲線下の面積 (area under the curve, AUC) AUC^{ROC} を分類器の性能評価に用いる．

ROC 曲線 [12] は，true-positive fraction と false-positive fraction をプロットしたものであり，sensitivity と specificity のトレードオフを表すものである．そして，AUC^{ROC} は ROC 曲線の下側の面積であり，ROC 曲線の良さを評価する尺度の一つである．AUC^{ROC} は 0 から 1 までの値をとり，完全な分類が可能な場合は 1 となり，ランダムな分類の場合は 0.5 となる．

本実験では，注目するクラスを正例，それ以外のクラスを負例として扱い，各クラスの AUC^{ROC} を計算した．さらに，各クラスの AUC^{ROC} を総合的に評価するために，AUC^{ROC} がしきい値以上であるクラスの割合を計算し，しきい値を変化させて，横軸に AUC^{ROC}，縦軸にクラスの割合をとってプロットした．その結果を**図 4** に示す．また，図 4 の曲線を定量的に評価するための曲線下の面積と対応する予測時間を**表 2** に示す．ただし，曲線下の面積は 0 から 1 までの値をとる．

図 4，表 2 に示すように，HNB の予測精度はナイーブベイズの予測精度を上回り，HNB はナイーブベイズの予測精度を改善していることがわかる．HNB の予測精度は SVM の予測精度よりは劣るが，ナイーブベイズと SVM の中間程度の予測精度が得られている．また，HNB の高速化に向けた改良の結果として，HNB (speed-up) は HNB の予測精度を低下させることなく約 5.5 倍の高速化を達成していることがわかる．結果として，HNB を用いることにより，ナイーブベイズに比べると予測時間が長くなるが，SVM と比べると非常に短い予測時間で SVM とナイーブベイズの中間の予測精度が得られる．しかし，NBH に関しては，今回使用したデータセットにおいてはナイーブベイズの予測精度を上回ることができなかった．

4.4 HNB の高速化

本研究では，3.2.1 で述べたように条件付き相互情報量にしきい値 t を導入することにより HNB の改良を行った．表 2 に示すように従来の HNB の予測精度をできるだけ低下させないようなしきい値 t を選択したが，**表 3** にしきい値 t を変化させたときの予測精度と予測時間を示す．ただし，従来の HNB の予測時間を短縮し，かつ，式 (8) の $p(a_i|a_{hp_i}, c)$ を 0 としないしきい値を選択した場合の結果を示す．また，HNB($t = 0$) は従来の HNB を示す．

結果として，条件付き相互情報量のしきい値 t を大きくするつれて予測時間は減少し，従来の HNB の予測精度を低下させることなく高速化を達成できるしきい値 t が存在したことがわかる．また，この結果からしきい値 t の決定が非常に重要であると考えられる．

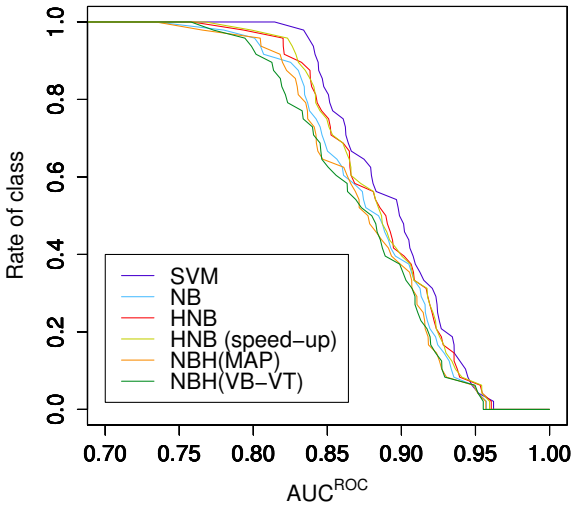

図 4 予測精度の比較
Fig. 4 Comparison of accuracy

表 2 曲線下の面積と予測時間
Table 2 Area under the curve and classification time

分類器	曲線下の面積	予測時間 (sec.)
SVM	0.893	21,457.4
NB	0.878	27.3
HNB	0.885	738.9
HNB (speed-up)	0.885	134.5
NBH (MAP)	0.875	57.4
NBH (VB-VT)	0.872	31,841.0

5. 結論

5.1 本研究の成果

本研究では，他分野でいくつかのデータセットに対しナイーブベイズより良い予測精度が得られることが報告されている拡張されたナイーブベイズである NBH と HNB の予測精度と予測時間を，UCI repository のデータセットを用いて検証した上でメタゲノム配列の系統分類に応用した．また，本研究の目的である高速かつ高精度な分類手法の開発のために，HNB による分類の高速化のための改良

表 3 条件付き相互情報量のしきい値 t を変化させたときの HNB の曲線下の面積と予測時間

Table 3 Area under the curve and classification time when threshold of conditional mutual information t is changed

分類器	曲線下の面積	予測時間 (sec.)
HNB ($t = 0$)	0.885	738.9
HNB ($t = 0.01$)	0.885	736.9
HNB ($t = 0.02$)	0.885	323.6
HNB ($t = 0.03$)	0.885	134.5
HNB ($t = 0.04$)	0.883	76.4
HNB ($t = 0.05$)	0.881	54.5
HNB ($t = 0.06$)	0.879	45.2

を行った．結果として，HNB を用いた系統分類手法の予測精度は SVM を用いた場合よりは劣るが，ナイーブベイズを用いる場合よりは向上していることが示された．したがって，HNB の導入により，SVM と比較すると非常に高速に分類することができ，SVM とナイーブベイズの中間程度の予測精度が得られる分類手法を開発することができた．また，本研究で用いたメタゲノム配列のデータセットにおいては，HNB の高速化のための改良により，予測精度を低下させることなく約 5.5 倍の高速化を達成した．

5.2 今後の課題

本研究では，DNA 断片配列から得られる特徴として 3-mer (3 塩基の部分配列) の出現頻度を分類器の属性に用いているが，先行研究の SVM を用いた Phylopythia [3] では 5-mer の出現頻度を属性とした場合に最も良い性能が得られている．また，ナイーブベイズを用いた系統分類手法 [4] では 12-mer から 15-mer の出現頻度を属性として用いた場合に良い性能が得られている．このため，我々も属性として適切な k-mer の k を選択することにより予測精度の向上が見込まれる．しかし，DNA 配列は A，T，G，C の 4 文字で表現されるため k-mer の出現頻度を属性とする場合，属性数は 4^k となり，k の値を大きくするにつれて属性数が非常に大きくなる．特に HNB のような分類器においては属性数が予測時間に大きく影響するため，選択する k の値によっては特徴選択により属性数を小さくする必要があると考えられる．

我々は，より高速な分類に向けて条件付き相互情報量のしきい値を導入することにより HNB の高速化を達成したが，現在はしきい値を cross-validation により決定している．このため，今後の課題として予測精度を低下させることなく高速化を達成するしきい値の決定方法を確立することが挙げられる．

また，今後 DNA シーケンサの更なる改善により今以上に短時間で大量のデータが出力されるようになり，現在よりもさらに高速な分類手法が求められると考えられる．

謝辞 本研究に関して有益な御助言を頂いた日本電信電話株式会社 NTT コミュニケーション科学基礎研究所の石畠正和氏に感謝の意を表する．

本研究は，文部科学省 博士課程教育リーディングプログラム「情報生命博士教育院」の支援を受けて行われたものである．

参考文献

[1] Huson, D. H., Auch, A. F., Qi, J. and Schuster, S. C.: MEGAN analysis of metagenomic data., *Genome Research*, Vol. 17, No. 3, pp. 377–386 (2007).
[2] Altschul, S. F., Gish, W., Miller, W., Myers, E. W. and Lipman, D. J.: Basic local alignment search tool., *Journal of Molecular Biology*, Vol. 215, No. 3, pp. 403–410 (1990).
[3] McHardy, A. C., Martín, H. G., Tsirigos, A., Hugenholtz, P. and Rigoutsos, I.: Accurate phylogenetic classification of variable-length DNA fragments., *Nature Methods*, Vol. 4, No. 1, pp. 63–72 (2007).
[4] Rosen, G., Garbarine, E., Caseiro, D., Polikar, R. and Sokhansanj, B.: Metagenome fragment classification using N-mer frequency profiles., *Advances in Bioinformatics*, Vol. 2008, p. 205969 (2008).
[5] Sato, T.: A General MCMC Method for Bayesian Inference in Logic-Based Probabilistic Modeling, *Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence*, pp. 1472–1477 (2011).
[6] Jiang, L., Zhang, H. and Cai, Z.: A novel Bayes model: Hidden Naive Bayes, *IEEE Transactions on Knowledege and Data Engineering*, Vol. 21, No. 10, pp. 1361–1371 (2009).
[7] Bache, K. and Lichman, M.: UCI Machine Learning Repository, `http://archive.ics.uci.edu/ml`.
[8] Kohavi, R.: Scaling up the accuracy of Naive-Bayes classifiers: A decision-tree hybrid, *In Proceedings of Second International Conferencde Knowledge Discovery and Data Mining*, pp. 202–207 (1996).
[9] Kurihara, K. and Welling, M.: Bayesian k-Means as a "Maximization-Expectation" Algorithm, *Neural Computation*, Vol. 21, pp. 1145–1172 (2009).
[10] Chang, C.-C. and Lin, C.-J.: LIBSVM: A library for support vector machines, *ACM Transactions on Intelligent Systems and Technology*, Vol. 2, No. 27, pp. 1–27 (2011).
[11] Richter, D. C., Ott, F., Auch, A. F., Schmid, R. and Huson, D. H.: MetaSim - A sequence simulator for genomics and metagenomics, *PLoS ONE*, Vol. 3, No. 10, p. e3373 (2008).
[12] Zweig, M. H. and Campbell, G.: Receiver-operating characteristic (ROC) plots: A fundamental evaluation tool in clinical medicine, *Clinical Chemistry*, Vol. 39, pp. 561–577 (1993).