平成 24 年度

学士学位論文

SPAM メールの判別に適した機械学習

Performance Comparison of Machine Learning Algorithms for SPAM Discrimination

1130378　　藤森 夏輝

指導教員　　吉田 真一

2013 年 3 月 9 日

高知工科大学 情報学群

要 旨

SPAM メールの判別に適した機械学習

藤森 夏輝

現在，SPAM メールか否かを判別する手法として，機械学習の一つであるナイーブベイズ分類器（ベイジアンネットワーク）があり，迷惑メールフィルタ製品として広く用いられている．しかし，その他の多くの機械学習手法を SPAM メールの特徴量に適用し，定量的に比較された研究はあまりない．そこで本研究では，SPAM メールの判別に，ナイーブベイズ分類器，ニューラルネットワーク，サポートベクターマシン（SVM），バギング，AdaBoost，RandomForest をそれぞれ適用し，英語と日本語の SPAM メール判別を行い，各手法の学習性能を明らかにする．英文 SPAM メール判別のデータには，UCI Machine Learning Repository 提供のデータセット「Spambase」を用いる．総数 4601 通のうち，ランダムに選出された 500 ～ 4000 通をそれぞれ訓練データとし，それぞれの残りをテストデータとして判別を行う．日本語 SPAM メール判別のデータセットは，独自に作成したコーパスを用いる．全 1400 通のうち，ランダムに選出された 1000 通を訓練データとし，残りをテストデータとして判別を行う．英語 SPAM メールにおいて同様の条件のもとで判別を行い，判別結果を比較する．結果として，英語 SPAM メール判別では，Bayes モデル以外の 5 手法が 90.6 ～ 94.9 %の判別率となることを示す．また，日本語の SPAM メールの判別率は，Bayes モデルが 42.8 %，そのほかの手法は 77.0 ～ 79.5 %の判別率となることを示す．

キーワード　SPAM，機械学習，ナイーブベイズ分類器，ニューラルネットワーク，サポートベクターマシン，バギング，AdaBoost，Random Forest

Abstract

Performance Comparison of Machine Learning Algorithms for SPAM Discrimination

Fujimori Natsuki

Many other machine learning techniques are able to applied for classification, and quantitative comparison is required. In this research, Naive Bayes Classifier, Neural Network, Support Vector Machine (SVM), Bagging, AdaBoost, and Random Forest are applied to classify e-mail written in both English and Japanese in order to filter SPAM mail out. Those algorithms are compared with each other from a viewpoint of classification precision. For English e-mail classification, the dataset "Spambase" of UCI Machine Learning Repository is used. Total number of the data is 4601, and training data is from 500 to 4000, which are randomly selected, and the rest are the test data. For Japanese e-mail classification, original corpus is created and used. Total number of the data is 1400 and training data are randomly selected to 1000. SPAM email in English discrimination is performed under the same conditions to compare the result of precision. As a result, for English SPAM distinction, all algorithms except Naive Bayes Classifier achieves the precision exceeding 90 %. Moreover, for Japanese SPAM, Naive Bayes Classifier became a distinction rate of 42.8 %, and 77-79 % abtained by other algorithms.

key words SPAM, Machine Learning, Naive Bayes Classifier, Neural Network, Support Vector Machine, Bagging, AdaBoost, Random Forest

目次

第 1 章　はじめに　1

第 2 章　採用手法　3

2.1　ナイーブベイズ分類器 3

2.2　ニューラルネットワーク 5

2.3　サポートベクターマシン 6

2.4　バギング (Bagging) 7

2.5　AdaBoost 8

2.6　Random Forest 9

2.7　交差検定法 10

第 3 章　SPAM メール判別実験　12

3.1　実験環境 12

3.2　英文コーパスを用いた SPAM メール判別実験 12

3.3　日本語コーパスを用いた SPAM メール判別実験 13

3.3.1　コーパスの作成 13

3.3.2　実験方法 16

第 4 章　実験結果および考察　17

4.1　英文コーパスを用いた SPAM メール判別実験 17

4.1.1　実験結果 17

各手法の判別実験 18

SVM のカーネル関数を用いた判別実験 19

4.1.2　考察 20

各手法の判別実験 20

SVM のカーネル関数を用いた判別実験 22
4.2 日本語コーパスを用いた SPAM メール判別実験 22
4.2.1 実験結果 22
4.2.2 考察 23

第 5 章 おわりに **25**

謝辞 **26**

参考文献 **28**

付録 A 英文コーパスにおける判別実験結果のグラフ拡大図 **29**

図目次

2.1 単一中間層ニューラルネットワーク 5
2.2 サポートベクターマシンの分離超平面 6
2.3 バギングの概略図 7
2.4 AdaBoost のアルゴリズムの簡略図 8
2.5 Random Forest の構造 10
2.6 8 分割交差検定法の概略図 11

3.1 日本語コーパス作成の流れ 13
3.2 作成した日本語コーパスの一部 15

4.1 各手法の SPAM 判別結果 18
4.2 SVM における 8 種類のカーネル関数を用いた SPAM 判別結果の表 19
4.3 SVM における 8 種類のカーネル関数を用いた SPAM 判別結果のグラフ . . 19
4.4 訓練データ数 3500 ～ 4000 のときの SPAM 判別の結果の表 21
4.5 訓練データ数 3500 ～ 4000 のときの SPAM 判別の結果のグラフ 21
4.6 日本語コーパスと英語コーパスにおける訓練データ数 1000 のときの SPAM 判別の結果のグラフ 24

A.1 図 4.1 の拡大図 29
A.2 図 4.3 の拡大図 30
A.3 図 4.5 の拡大図 30

表目次

2.1 100 通のメールに出現した単語の一例 4
2.2 求めた確率の一覧 4

3.1 実験に用いたハードウェア・ソフトウェア 12
3.2 日本語コーパスに用いた単語一覧 14

4.1 6 種類の学習手法による SPAM 判別の結果 17
4.2 日本語コーパスと英語コーパスにおける訓練データ数 1000 のときの SPAM 判別の結果の表 23

第 1 章

はじめに

携帯電話端末やインターネットにおける主要なコミュニケーション手段の一つに電子メールが挙げられる．電子メールがコミュニケーションツールとして一般的になった一方で，利用者が希望していない情報やコンピュータ・ウィルスを添付したメールを無作為に送りつける，いわゆる SPAM メールの増加が深刻な問題になっている．文献 [1] では全世界のメールトラフィックに占める SPAM メールの割合が 2010 年には 92.51 %を超えたという報告が示されている．

受信メールから SPAM メールを除外する SPAM メールフィルタでは，アドレスやドメインなどの発信元情報，ヘッダやメール本文の出現単語などの特徴を学習し，SPAM メールであるか否かの判別を行なっている．この SPAM メールフィルタに利用される技術が機械学習アルゴリズムである．機械学習アルゴリズムは多くの手法が考案されており，中でもベイジアンフィルタとして知られるナイーブベイズ分類器が SPAM メールフィルタとして広く利用されている．これら機械学習アルゴリズムの間で，個々に分類問題に対する性能評価を行った研究は過去にあるが，多くの機械学習手法に対して定量的に比較を行ったものはない．

そこで本研究では，いくつかの機械学習手法について SPAM メール判別を対象として，同じデータセットを用いて判別を行い，その判別性能の定量的比較を行うことを目的とする．英語で記述された SPAM メールのコーパス（以下英文コーパス）と日本語で記述された SPAM メールのコーパス（以下日本語コーパス）を用いて，以下の 6 手法において SPAM メール判別を行う．

- ナイーブベイズ分類器（以下 Bayes）
- ニューラルネットワーク（以下 NN）
- サポートベクターマシン（以下 SVM）
- バギング
- AdaBoost
- Random Forest（以下 RF）

英文コーパスは，Hewlett-Packard Labs にて Mark Hopkins らが作成した，UCI Machine Learning Repository が無償で提供している「Spambase Data Set」と，独自に作成した日本語 SPAM メールデータセットを利用して実際に機械学習を行う．作成した日本語メール用データセットは，独自に入手した 22317 通の日本語 SPAM メールおよび非 SPAM メール（以下 HAM メール）から 1400 通を選出し，形態素解析を行いその出現頻度に TF 値と IDF 値をかけたもので構成されている．判別実験では，英文コーパスについては訓練データ数を 500 ～ 4000 まで 500 刻みに増加させ，学習を繰り返しその判別性能を比較する．また，日本語コーパスについては訓練データ数 1000 の場合について各手法の判別率を，英文コーパスの同条件下における判別率と比較する．

英文コーパスにおける SPAM 判別実験を行い，Bayes モデルを除く 5 手法について，判別率が常時 90 %を超えることを示す．また，日本語コーパスにおける SPAM 判別実験では，同じ 5 手法の判別率が 75 ～ 79 %，Bayes モデルについては 42.2 %となり，英文コーパスの同条件下における判別率と比べ全体的に 10 ～ 15 %低い結果となる．

本論文では，第 2 章で今回実験を行うにあたって採用した 6 手法についてのアルゴリズムの説明を行い，第 3 章で SPAM メール判別実験の詳細，および日本語コーパスの生成方法について述べ，第 4 章で SPAM メール判別実験の結果と考察を述べる．

第 2 章

採用手法

本章では，SPAM メール判別実験に採用した 6 つの機械学習手法および 1 評価手法についての紹介とアルゴリズムの説明を行う．

2.1 ナイーブベイズ分類器

ナイーブベイズ分類器（Naive Bayes classifier）とは，ベイズ推定に基づく機械学習手法である．ベイズ推定に用いられるベイズの定理は，以下の数式で表される．

$$P(B|A) = \frac{P(A|B)P(B)}{P(A)} \tag{2.1}$$

ここで，$P(A), P(B)$ は事象 A（もしくは B）が発生する確率，$P(B|A)$ は事象 A が発生した後の事象 B が発生する確率を表す．

例として，100 通のメールの SPAM 判別を行うとする．100 通のうち，40 通が SPAM，60 通が HAM メールとする．ここで，B_S を SPAM メールが出現する事象，B_H を HAM メールが出現する事象とすると，$P(B_S)$ は SPAM メールの出現確率，$P(B_H)$ は HAM メールの出現確率である．すなわち，$P(B_S) = 40/100 = 0.4$，$P(B_H) = 60/100 = 0.6$ となる．また，本研究では単語の出現頻度を元に SPAM 判別を行うので，この例でも同様に単語の出現頻度を用いる．表 4.1 は，100 通のメール中に出現した単語とその頻度の一例である．このデータをもとに，各単語の出現確率である $P(A), P(A|B)$ を求める．ここでは,「メール」という単語についての各確率を示す．ここで事象 $A =$ ”メール” とは,「メール」という単語が出現することを示す．よって，$P(A =$ ”メール”$) = (5+11)/100 = 0.16$ である．$P(A =$ ”メール”$|B_S)$ は SPAM メールの中で「メール」という単語が発生する確率である．故に，

term	SPAM	HAM
メール	5	11
よろしく	3	14
配信	13	2
淫乱	20	0
・	・	・
・	・	・
・	・	・

表 2.1　100 通のメールに出現した単語の一例

$P(A = \text{"メール"}|B_S) = 5/40 = 0.125$”，$P(A = \text{"メール"}|B_H) = 11/60 = 0.1833$ となる．表 2.2 に，ここまでに求めた各確率を示す．以上の確率から，単語「メール」の SPAM

	SPAM	HAM
$P(B)$	0.4	0.6
単語「メール」に関数する確率		
$P(A)$	0.16	
$P(A\|B)$	0.125	0.1833

表 2.2　求めた確率の一覧

らしさ，HAM らしさを推定する．まず SPAM らしさの推定である，$P(SPAM|$ メール$)$ は，$0.125 \times 0.4/0.16 = 0.3125$ となる．続いて，HAM らしさの推定である，$P(HAM|$ メール$)$ は，$0.1833 \times 0.6/0.16 = 0.687$ となり，$P(SPAM|$ メール$) > P(HAM|$ メール$)$ で「メール」は HAM と分類される．このような処理を，出現する単語数分（利用するデータセットに含まれる単語数分）実行する．

以上のように，アルゴリズムが単純（ナイーブ）であることからナイーブベイズ分類器と呼ばれる．また，計算式が簡略化されているため計算コストも低く，Mozilla Thunderbird

のようなメーラーの迷惑メールフィルタにも用いられている．さらに，ナイーブベイズ分類器は，訓練データを多くすればするほど汎化性能が向上するという特徴を持っている．

2.2　ニューラルネットワーク

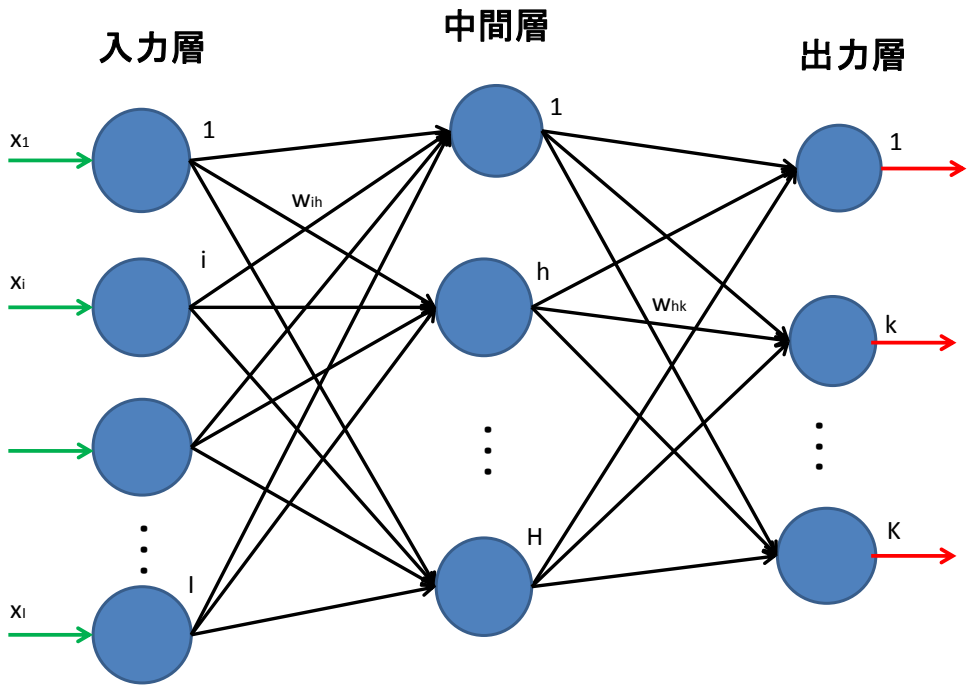

図 2.1　単一中間層ニューラルネットワーク

ニューラルネットワーク（Neural Network）は，David E. Rumelhart が 1986 年に提案した，バックプロパゲーション学習を用いている，人間の神経回路を模した機械学習手法の一つである．ニューラルネットワークには階層型モデルと非階層型モデルの 2 種類のモデルが存在する．今回実験で利用したのは階層型モデルである．階層型モデルはニューラルネットワークの中でも最も多く利用されているモデルである．図 2.1 は単一中間層ニューラルネットワークと呼ばれ、次式により定式化される [3].

$$y_k = \phi_0(\alpha_k + \Sigma_h w_{hk}\phi_h(\alpha_h + \Sigma_i w_{ih}x_i)) \tag{2.2}$$

ここで，ϕ は伝達関数であり，通常はシグモイド関数が用いられる．α は定数である．また、

結合の重みを定める学習は，以下のステップとなる．

1. 訓練データをニューラルネットワークに入力する．最初の段階では，結合の重みは小さなランダム値が付与される．この重みを用いて，1 回目の出力を行う．
2. 出力結果と学習データを比較し，次式により重みの更新を行う．

$$w(j+1) = w(j) + \eta \times \delta \times R \tag{2.3}$$

ここで，$w(j+1)$ とは $j+1$ 回目の重み，$w(j)$ とは j 回目の重み，η は学習定数，δ は各層でそれぞれ値の異なる，出力結果と学習用データとの差の関数（出力誤差），R は出力結果を示している．

3. 最適解が得られるまで，2. を繰り返す．

2.3　サポートベクターマシン

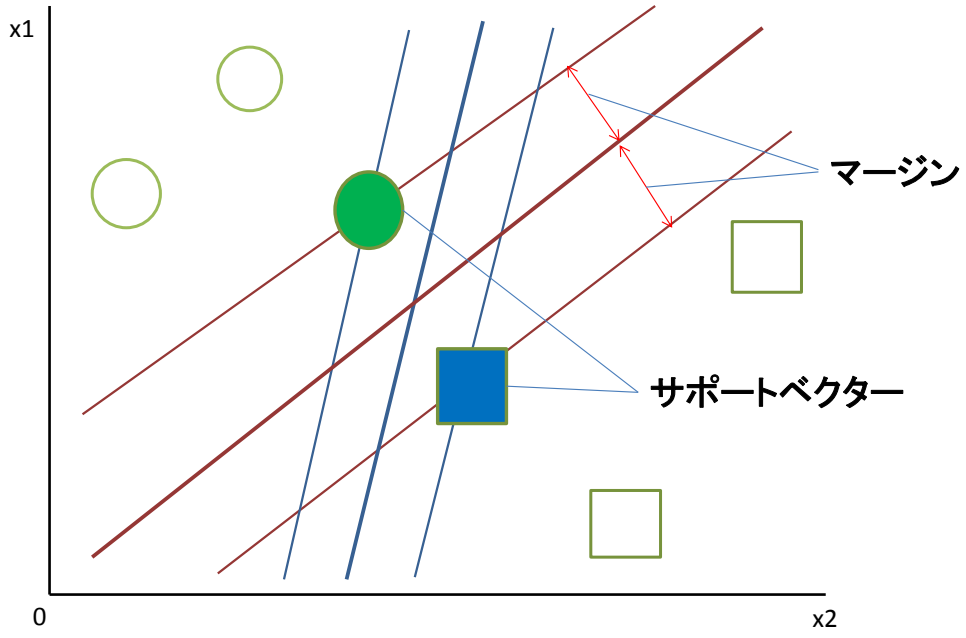

図 2.2　サポートベクターマシンの分離超平面

サポートベクターマシン（Support Vector Machine, SVM）は，Vladimir N.Vapnik らが 1992 年に提案した機械学習手法である．SVM は，高次元特徴空間において線形関数の仮設空間を用いる学習システム [2] である．学習データを仮設空間上にマッピングし，両クラスのデータ点間の距離が最大となる分離超平面を求める．分離直線に一番近い各クラスのデータ点をサポートベクターと呼び，両サポートベクター間の距離（マージン）を最大化することで，未知の学習データに対して高い汎化能力を有するという特徴を持つ．図 2.2 に，2 次元空間における線形分離可能なデータのマージン最大化の概略図を示す．SVM は，元々線形分離可能な問題に対する分類器として開発されたが，カーネル関数を用いることで非線形分離問題に対しても有用な学習手法となった．

2.4　バギング (Bagging)

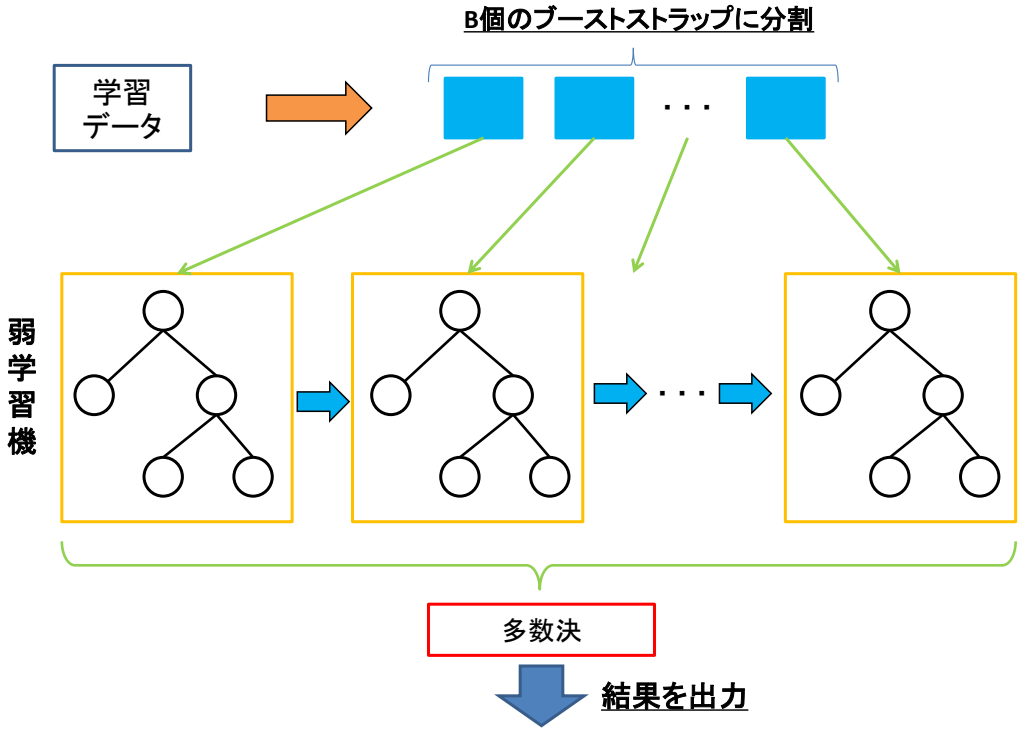

図 2.3　バギングの概略図

バギング（Bagging）は，Leo Breiman が 1996 年に提案した集団学習の手法である．集

団学習とは，低精度な複数の学習結果を組み合わせ，精度の向上を図る学習法の総称である．バギングでは，bootstrap（ブートストラップ）と呼ばれる複数の学習データを，与えられた元のデータセットからサンプリング法により複数作成し，以下の手順にて学習を行い精度を向上させる．図 2.3 は以下に示したアルゴリズムの概要を図示したものである．

1. n 個の個体から構成される学習データを用意する．
2. 学習データから復元抽出法を用いて m 回抽出し，ブートストラップを作成する．
3. 弱学習機モデル h を作成し，判別を行う．
4. 1. ～ 3. を B 回繰り返し，判別モデルを B 個 $\{h_i|i = 1, 2, ..., B\}$ 作成する．
5. 以下の数式を用いて 4. で得られた判別モデルの多数決をとり，強学習機とする．

   $H(x) = \arg\max |\{i|h_i = y\}|$

2.5　AdaBoost

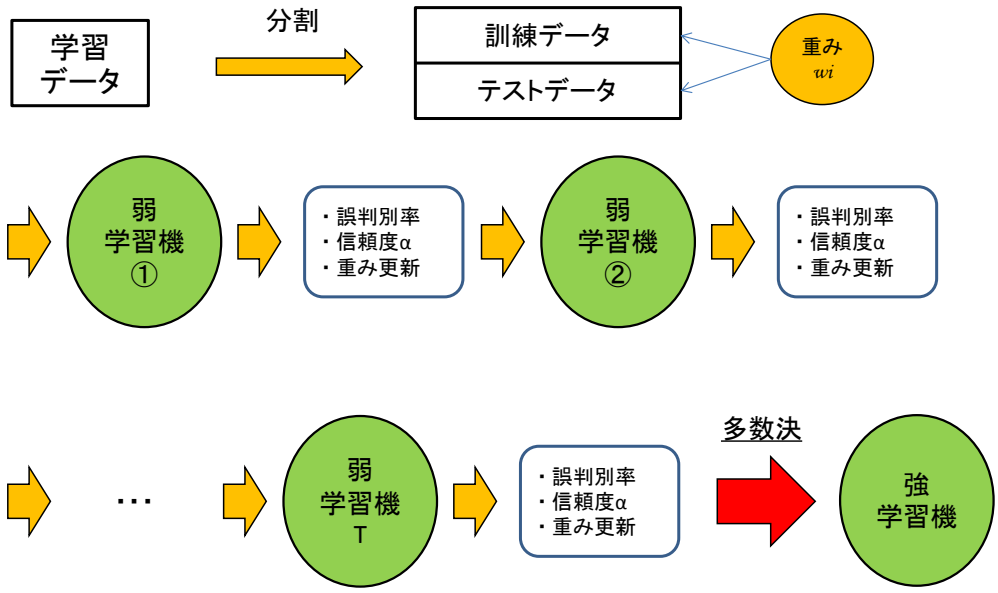

図 2.4　AdaBoost のアルゴリズムの簡略図

AdaBoost は，Yoav Freund と Robert Schapire が 1996 年に提案した集団学習手法である．学習データを複数に分割し，それぞれを弱学習機に学習させその多数決をとるという点では前節で説明したバギングと同じである．AdaBoost では，学習データに重みを付加することで学習精度の向上を図っている．この重みは，初期値が 1/（データ数） となるが，2 番目以降の弱学習機に対して，前の学習結果の誤判別率が高ければ低く，誤判別率が低ければ高く設定し，誤判別しやすい特異データに対してマーキングを行い，誤判別を減少させるという特徴がある．図 2.4 は AdaBoost の概要を図示化したものである．以下に AdaBoost のアルゴリズムの概要を示す.

1. N 個の学習データを用意する.
2. T 個の弱学習機を生成すると仮定する.
3. 重みを $w_{ti} = 1/N$ とする.（初期化）
4. 重みを用いて学習を行い，弱学習機 t を構成する.
5. 弱学習機の誤判別率を求める．誤判別率は,
6. 誤判別率を用いて弱学習機 h の信頼度 α を求める
7. 重み $w_{(h+1)i}$ を更新する.
8. 3. ～ 7. を T 回繰り返す.
9. 全弱学習機を信頼度 α で重み付けし，多数決をとって強学習機とする.

2.6 Random Forest

Random Forest は，バギングを提案した Leo Breiman によって 2001 年に提案された集団学習手法である．図 2.5 は Random Forest の概要を図式化したものである．まず入力データから 組のブートストラップサンプルを生成する，次に，生成したブートストラップサンプルを用いて木構造の弱学習機を生成する．この時、分岐ノードにはランダムサンプリングされた変数を用いる．最後に，生成した弱学習機による学習結果の多数決を取り、強学習機を構築する．この強学習機にテストデータを通すことで，分類を行うことができる.

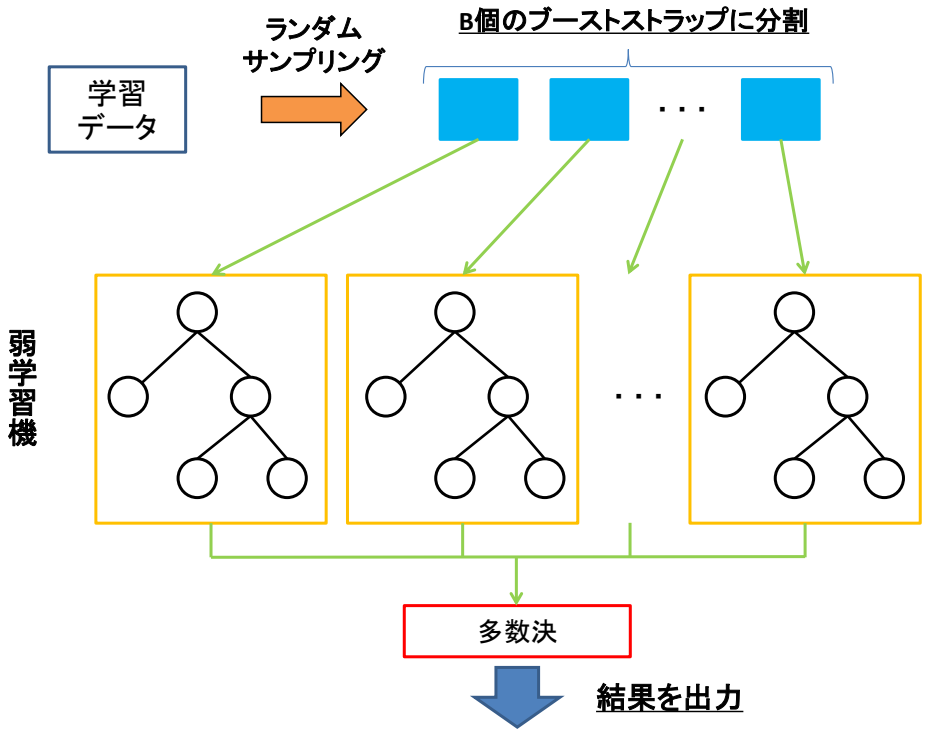

図 2.5　Random Forest の構造

Random Forest は学習精度が高く，次元数の大きいデータに対して頑健であるという特徴を持っているため，今回の実験で用いることを決めた．以下に Random Forest のアルゴリズムの概要を示す．

1. 学習データから B 組のデータストラップを生成する．
2. 各ブートストラップとランダムサンプリングされた変数を用いて弱学習機を生成する．
3. 各弱学習機を生成後，各々の学習結果の多数決を取り、強学習機とする．

2.7　交差検定法

機械学習機で分類した訓練データは，評価データを用いて正しく分類できているかどうかを判定し，その結果をもとにテストデータを分類する．しかし，評価データが存在しないようなデータの分類には，交差検定（Cross Validation）法を用いる．図 2.6 は，8 分割交差

2.7　交差検定法

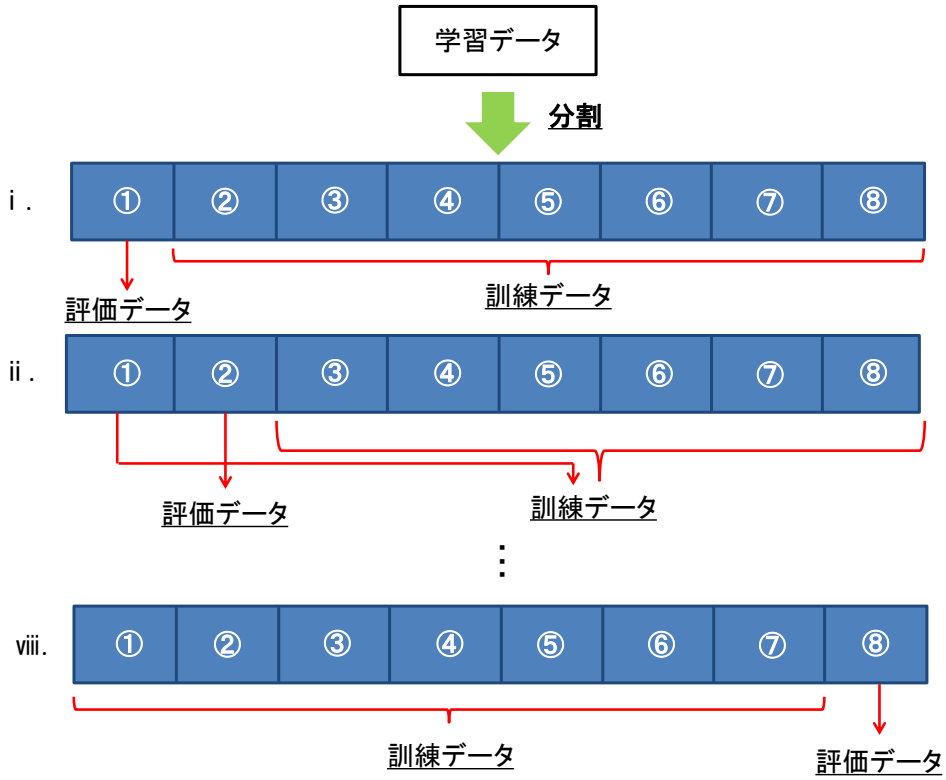

図 2.6　8 分割交差検定法の概略図

検定法の模式図である．以下に交差検定法のアルゴリズムの概要を示す．

1. データセットを a 個に分割する．
2. 1 つ目のデータを評価データ，残りのデータを訓練データとし，分類の評価を行う．
3. 2 つ目のデータを評価データ，残りを訓練データとし，評価を行う．このとき，1. で用いた評価データは訓練データとなる．
4. 以下同様に，a 個目までのデータを評価データとして評価するまで繰り返す．
5. 4. までの評価で得た正解率などの評価における指標値を平均し，そのデータセットに対する最終的な指標値とする．

第 3 章

SPAM メール判別実験

3.1 実験環境

本実験を行うにあたって用意した環境は表 3.1 のとおりである．

OS	Windows 7 Enterprise
CPU	Intel(R) Core(TM) i5-2400S CPU @ 2.50GHz
メモリ	4.00GB
利用ソフトウェア	R x64 2.15.2 （統計解析ソフト） MeCab （形態素解析ソフト）

表 3.1　実験に用いたハードウェア・ソフトウェア

3.2 英文コーパスを用いた SPAM メール判別実験

英文コーパスを利用した英語 SPAM メールの判別実験について説明を行う．利用するデータセットは，UCI Machine Learning Repository 提供のデータセット「Spambase Data Set」を利用する．このデータセットには 4601 通のメールが格納されており，それぞれに単語出現頻度，記号使用頻度，大文字平均値，最長文字数，文字総数，Spam or NonSpam 情報からなる 58 次元の特徴量が含まれている．4601 通のうち，1813 通が SPAM メール，2788 通が HAM メールとなっている．

本実験では，訓練データ数を 500 〜 4000 の間で，500 刻みに増加させた場合の学習性能

を比較するため，各手法で判別実験を行うものである．

3.3　日本語コーパスを用いた SPAM メール判別実験

3.3.1　コーパスの作成

本実験を行うにあたって，利用できる日本語コーパスが見つからなかったため，コーパスを独自に作成することとした．本節ではコーパス作成手順について説明する．本実験では，単語の使用頻度のみを特徴量として用いることとする．図 3.1 は以下の手順を図示したものである．

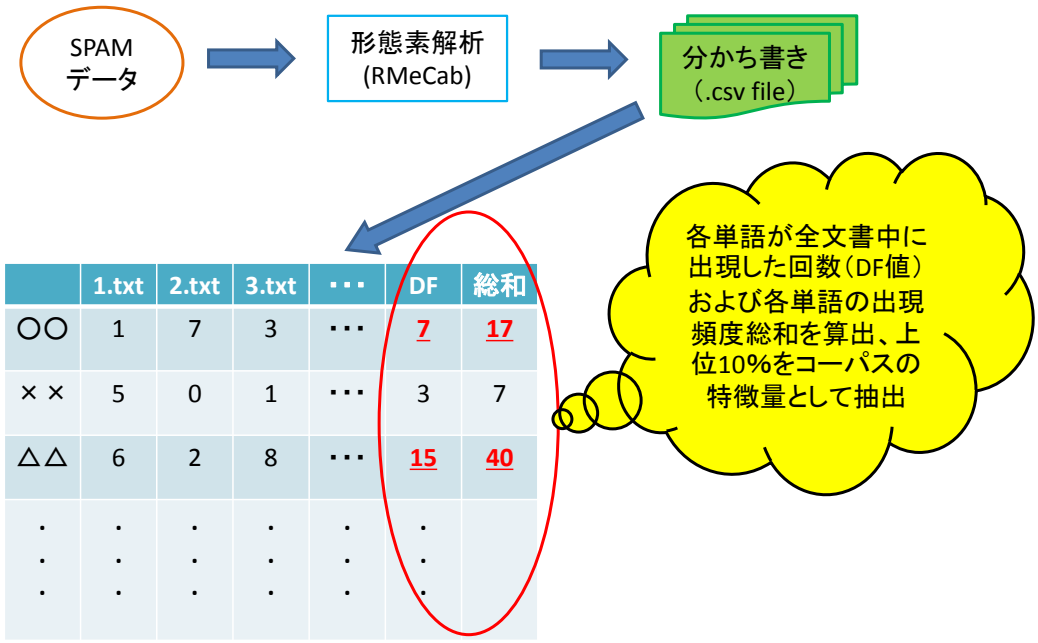

図 3.1　日本語コーパス作成の流れ

1. コーパスに実際に格納するメールとは別の SPAM メールを 100 通取得する．
2. 取得したメールを形態素解析し，出現頻度を算出する．
3. 形態素解析した結果から，日本語である単語の行を抽出する．

4. 抽出後，単語が全文書中において何文書出現しているかの頻度（Document Frequency, DF 値）を全単語算出する．
5. 各単語の出現頻度の総和を算出する．
6. 4. で算出した DF 値の上位 10 %および 5. で算出した頻度総和の上位 10 %をコーパスの単語とする．

また，表 3.2 は日本語コーパスの特徴として選んだ単語の一覧である．なお，コーパスには表 3.2 に示した単語のほか，SPAM か HAM かを示す ”type” がある．図 3.2 は実際に作成した日本語コーパスの一部を抜粋したものである．

サイト	配信
無料	方
様	言う
今回	見る
女性	人
登録	探す

表 3.2　日本語コーパスに用いた単語一覧

	サイト	無料	様	今回
1	0	0	0	0
10	0	0	0	0
100	15.72893026	2.104697379	11.35762558	0
101	0	4.209394757	0	9.473931188
102	0	0	0	0
103	0	0	0	0
104	2.621488377	6.314092136	0	4.736965594
105	0	0	0	0
106	0	0	0	0
107	2.621488377	4.209394757	0	0
108	2.621488377	4.209394757	0	0
109	0	0	0	0
11	0	4.209394757	0	0
110	0	0	0	0
111	0	2.104697379	0	0
112	2.621488377	0	3.785875195	0
113	0	0	0	0
114	0	0	0	0
115	2.621488377	0	0	4.736965594
116	0	4.209394757	0	0
117	0	4.209394757	0	0
118	0	0	0	0
119	0	0	0	0

図 3.2　作成した日本語コーパスの一部

3.3.2　実験方法

本実験を行うにあたって作成したコーパスには 1400 通分のメールが含まれており，そのうち 800 通が SPAM メール，残り 600 通が HAM メールとなっている．性能評価を行う目的で，訓練データを 1000 通として判別を行う．さらに，英文コーパスでも同様の条件で判別実験を行い，それぞれの性能の比較検証を行う．

第 4 章

実験結果および考察

本章では，第 3 章で説明した実験の結果を示し，その考察を述べる．

4.1 英文コーパスを用いた SPAM メール判別実験

4.1.1 実験結果

	Bayes	NN	SVM	バギング	AdaBoost	RF
500	0.6559376	0.919288	0.8973421	0.9051451	0.9053889	0.9256279
1000	0.6842544	0.9297417	0.9050264	0.9122466	0.9139128	0.9394613
1500	0.6875202	0.9245405	0.9129313	0.9068043	0.9090616	0.9422767
2000	0.6885813	0.9300269	0.9200308	0.9073433	0.9161861	0.9423299
2500	0.7029986	0.9324131	0.9281295	0.9033793	0.9219419	0.9433603
3000	0.7083073	0.9394129	0.9312929	0.9050593	0.9125547	0.9437851
3500	0.7111717	0.9445958	0.9300636	0.9064487	0.9218892	0.9491371
4000	0.7304493	0.9351082	0.9267887	0.9018303	0.9301165	0.9450915

表 4.1　6 種類の学習手法による SPAM 判別の結果

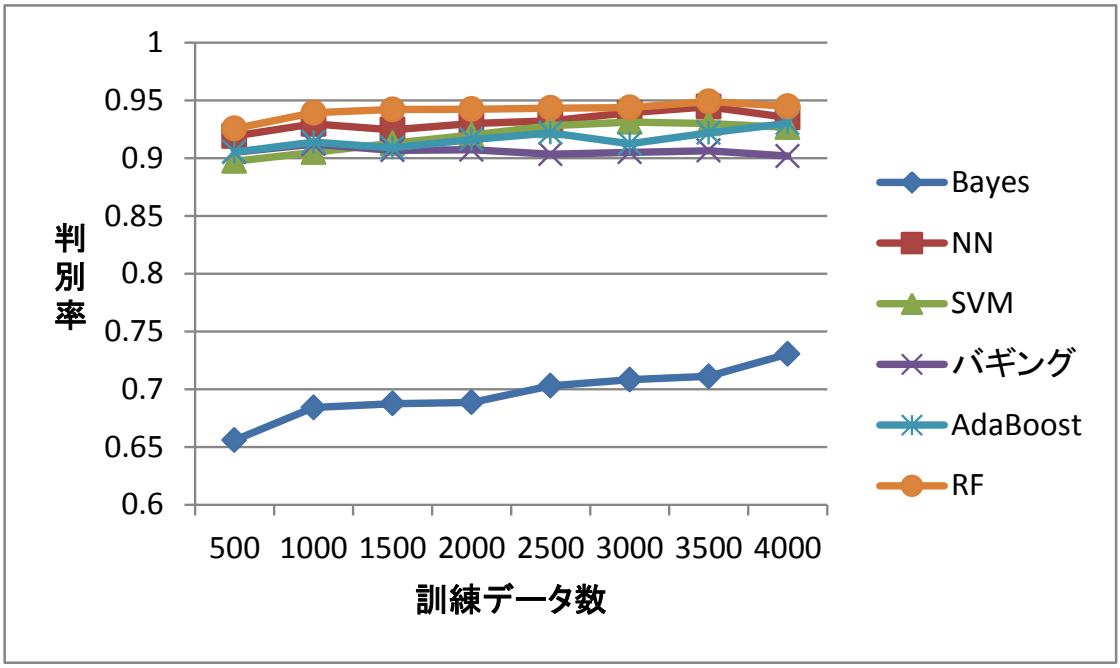

図 4.1　各手法の SPAM 判別結果

各手法の判別実験

表は訓練データ数を 500 ～ 4000 まで 500 刻みに増加させた際の各手法の判別率を示したものであり，図 4.1 は表をグラフ化したものである．表・図中の NN はニューラルネットワーク，RF は Random Forest を表し，表中の第 1 列には訓練データ数，第 1 行にはそれぞれの手法名を記載している．本節に提示する表の判別率は，有効数字 7 ケタを使用しているが，これは各手法において前後の訓練データ数との差が乏しく，より細かい値が結果に影響を及ぼしており，四捨五入により値を丸めるのは不適切であると考えたため，計算機の出力結果をそのまま記載している．

SVM のカーネル関数を用いた判別実験

	ガウシアン	線形	多項式	タンジェント	ラプラシアン	ベッセル	ANOVA	スプライン
500	0.891733723	0.9029505	0.86100951	0.799804926	0.907583516	0.545476713	0.911241161	0.844672031
1000	0.910024993	0.90863649	0.883365732	0.777006387	0.9130797	0.493474035	0.926409331	0.860038878
1500	0.921315705	0.917445985	0.882296034	0.790712673	0.922605611	0.546597872	0.930667527	0.862947436
2000	0.9242599	0.924644368	0.894655902	0.796232218	0.92272203	0.517877739	0.931180315	0.902729719
2500	0.926701571	0.924321752	0.899095669	0.74631128	0.925749643	0.540218943	0.938124703	0.910042837
3000	0.931917552	0.931292942	0.90131168	0.797001874	0.927545284	0.552779513	0.926296065	0.555902561
3500	0.934604905	0.924613987	0.905540418	0.792915531	0.920980926	0.522252498	0.933696639	0.821071753
4000	0.931780366	0.916805324	0.908485857	0.792013311	0.920133111	0.510815308	0.93344426	0.821071753

図 4.2　SVM における 8 種類のカーネル関数を用いた SPAM 判別結果の表

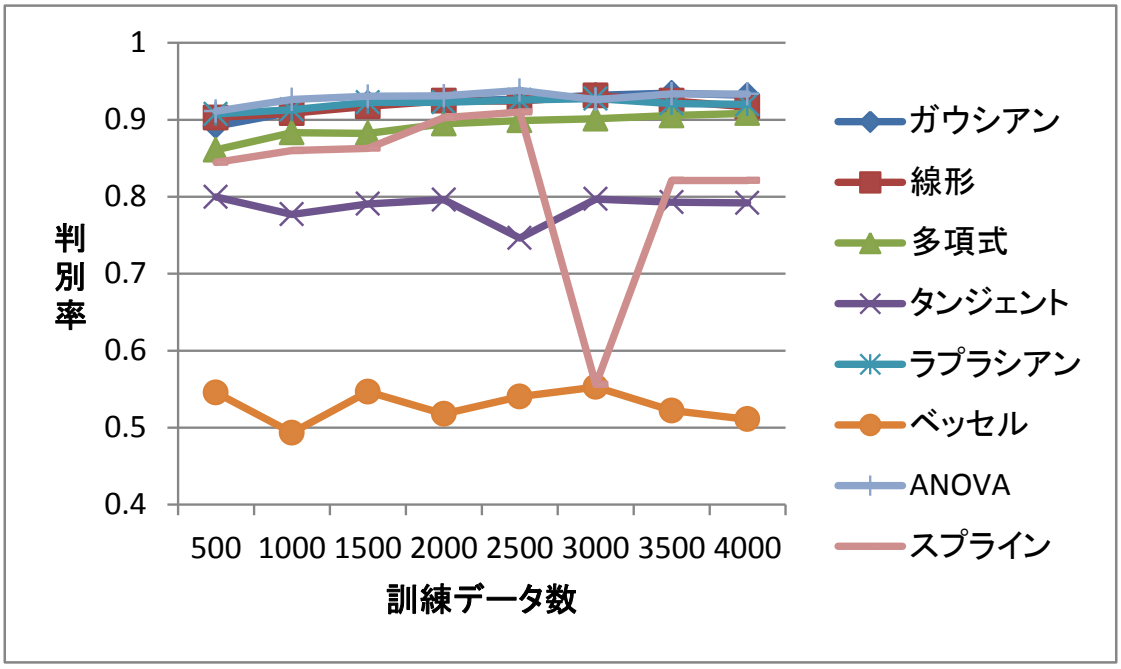

図 4.3　SVM における 8 種類のカーネル関数を用いた SPAM 判別結果のグラフ

図 4.2 は SVM について，8 種類のカーネル関数を用いて同様の実験を行った際の各関数の判別率を表で示したものであり，図 4.3 は表をグラフ化したものである．表中の第 1 列に

訓練データ数，第 1 行にカーネル関数名を記載している．

4.1.2　考察

各手法の判別実験

表や図から，ナイーブベイズ分類器の判別性能が他の手法に比べ 20 %も低い結果となった．それに対し，6 手法の中で最も判別性能の高い手法は Random Forest であった．ただし，ほかの手法が訓練データ数 3500 を境に下降傾向に転じているのに対し，ナイーブベイズ分類器は常時上昇傾向になっていることから，まだ性能向上の余地があると考えられる．性能向上を図る方法として，訓練データ数の増加が考えられるが，本実験に用いたデータセットが 4601 通分であるため，より多くのデータ数をもつ他のデータセットを利用した実験が望ましいと考えられる．

また，前述したように，ナイーブベイズ分類器を除く 5 手法の判別性能が，訓練データ数 3500 を境に下降傾向に転じた．このことから，このデータセットにおいて，教師データを学習しすぎることによって汎化性能を下げる過学習の状態が，3500 〜4000 の間に発生していると考えられる．以下に，この事象について行った追加実験の結果を示す．

図は訓練データ数 3500 〜 4000 の間において，50 刻みにデータ数を増加させてその判別率を比較した表である．図はこれをグラフ化したものである．SVM に関しては ガウシアン RBF カーネルを使用して判別を行った．

表及び図から，各手法において，最も高い判別率を記録した訓練データ数にばらつきが出ることがわかった．本実験においては，ナイーブベイズ分類器が訓練データ数 3950 の時に 74.5 %，ニューラルネットワークと Random Forest が訓練データ数 3800 の時にそれぞれ 95.0 % と 96.3 %，SVM が訓練データ数 3600 の時に 94.1 %，バギングと AdaBoost が訓練データ数 3500 の時にそれぞれ 92.2 % と 94.3 %という結果になった．また，追加実験前，ナイーブベイズ分類器はデータ数を増やすことで判別率が上昇するという予測を立てていたが，この結果では急激な上昇などは観測されず，80 %を超えることはなかった．この原

	Bayes	NN	SVM	Bagging	AdaBoost	RF
3500	0.713896458	0.930971844	0.933696639	0.921889192	0.942779292	0.959128065
3550	0.703139867	0.941960038	0.917221694	0.900095147	0.917221694	0.939105614
3600	0.732267732	0.946053946	0.941058941	0.911088911	0.93006993	0.947052947
3650	0.703470032	0.936908517	0.930599369	0.904311251	0.923238696	0.954784437
3700	0.720310766	0.933407325	0.937846837	0.905660377	0.92563818	0.955604883
3750	0.722679201	0.937720329	0.929494712	0.902467685	0.92479436	0.945945946
3800	0.696629213	0.950062422	0.937578027	0.913857678	0.936329588	0.962546816
3850	0.711051931	0.917443409	0.934753662	0.909454061	0.925432756	0.954727031
3900	0.697574893	0.932952924	0.92296719	0.910128388	0.915834522	0.938659058
3950	0.74500768	0.935483871	0.937019969	0.894009217	0.913978495	0.950844854
4000	0.727121464	0.936772047	0.935108153	0.91014975	0.913477537	0.951747088

図 4.4　訓練データ数 3500 ～ 4000 のときの SPAM 判別の結果の表

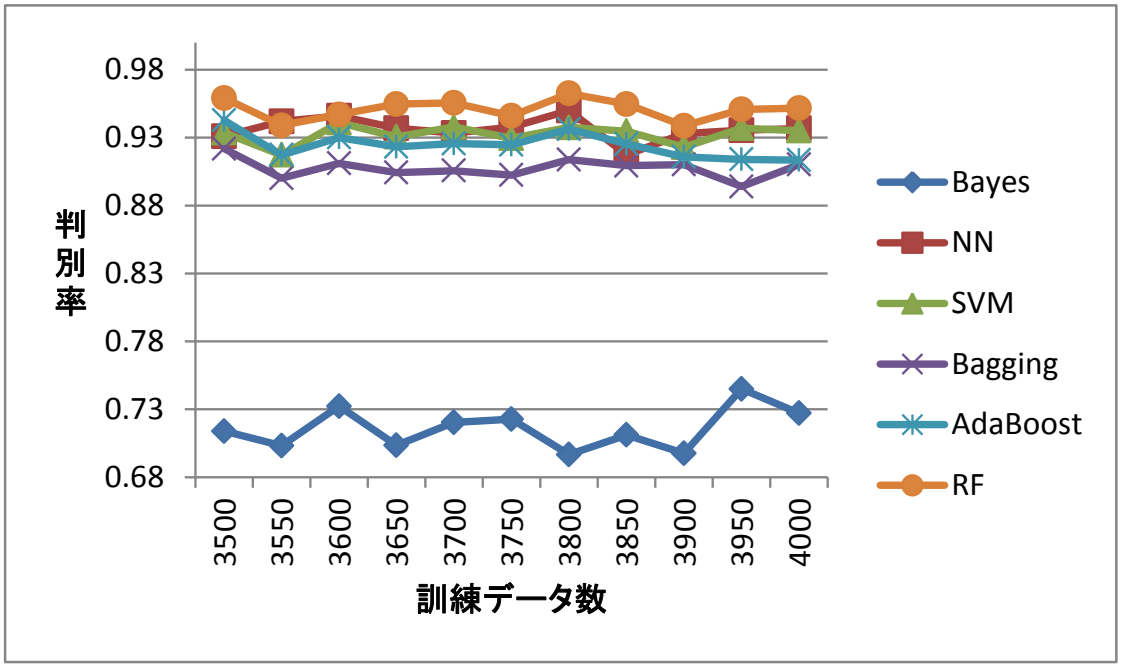

図 4.5　訓練データ数 3500 ～ 4000 のときの SPAM 判別の結果のグラフ

因として，コーパスの特徴数に応じて判別性能の限界があるのではないかと考えられる．

SVM のカーネル関数を用いた判別実験

表や図から，ベッセルカーネルを用いた判別は，精度が 50% 前後となり， SPAM 判別に適さないことがわかった．それに対し，8 種類のカーネル関数の中で最も判別性能の高い関数は ANOVA 関数であった．

また，こちらも前述した 6 手法における SPAM 判別実験で得た結果と同じように，判別率が 90% を超える関数において，訓練データ数 3500 の際に判別性能が最高値に達し，その後は下降傾向に転じ現象が見られた．

また，スプラインカーネル関数に関して，訓練データ数 2500 までは順調に精度を上げていたが，3000 で判別を行った際に 55% まで急落するという現象が発生した．原因として考えられることとして，R 内部の演算にエラーが生じている可能性が挙げられるが，このときエラーは発生していなかったので，他に原因があるように考えられるが，現在その原因は不明である．スプラインカーネルは最高 90% を超える関数であり，2500 以降も上昇することが考えられる．故に，この現象に関して原因を究明し精度の向上を目指すことは今後の重要な課題といえる．

4.2　日本語コーパスを用いた SPAM メール判別実験

4.2.1　実験結果

図 4.2 は 日本語コーパスと英語コーパスに関して，訓練データ数 1000 のときにパラメータを固定して SPAM 判別実験を行ったものの結果を表で示したものであり，図 4.6 は表をグラフ化して視覚的に比較したものである．表中の NN はニューラルネットワーク，RF は Random Forest を示し，第 1 列に機械学習の手法名，第 1 行にコーパスの種類を記載している．

	英文	日本語文
Bayes	0.684	0.422
NN	0.929	0.785
SVM	0.926	0.795
バギング	0.912	0.778
AdaBoost	0.913	0.77
RF	0.939	0.793

表 4.2　日本語コーパスと英語コーパスにおける訓練データ数 1000 のときの SPAM 判別の結果の表

4.2.2　考察

グラフによる比較結果から，全体的に英文コーパスの方が優れていることが示された．これは，コーパスに含まれている特徴量の差に原因があると考えられる．日本語コーパスは単語出現頻度のみを特徴量の項目として採用しているのに対し，英文コーパスは同じく単語出現頻度に加え，記号出現頻度，大文字平均値，最長文字数，総数を特徴項目として採用している．判別精度向上を目指すためにも，日本語コーパスの特徴量項目を再考することが今後の重要な課題である．

また，本実験の結果では，0.02% 差で SVM が最も判別性能の高い手法となった．ただし，学習量がまだまだ少なく，この先も上昇することが考えられるため，特徴量項目の再考に加え，データ数の増加も今後の課題といえる．

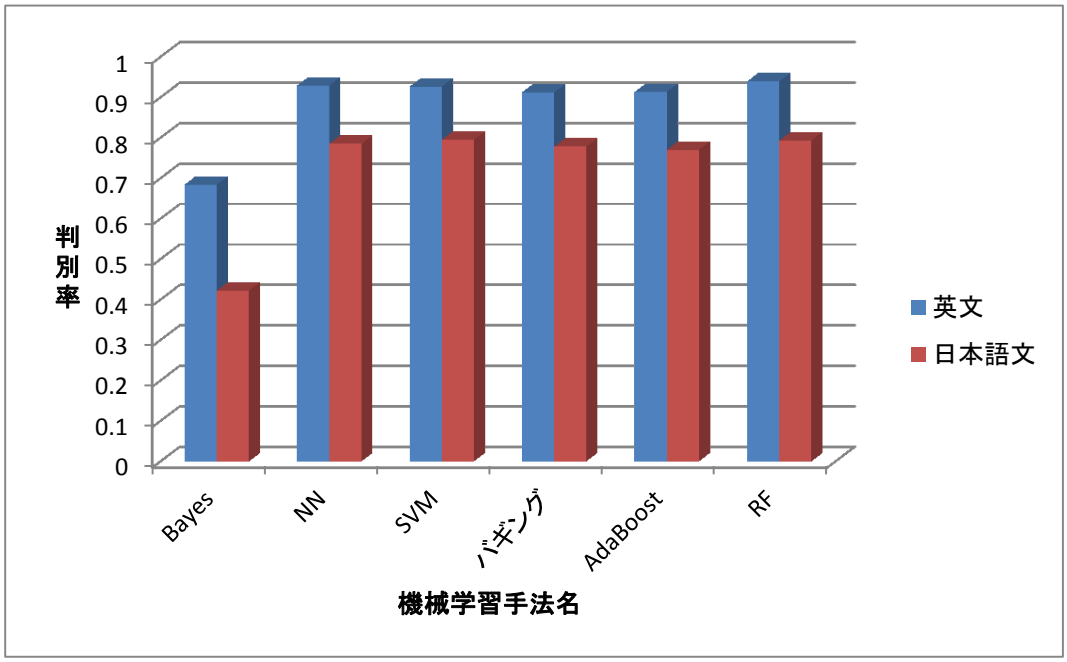

図 4.6　日本語コーパスと英語コーパスにおける訓練データ数 1000 のときの SPAM 判別の結果のグラフ

第 5 章

おわりに

本研究では，増加傾向にある SPAM メールを受信メールから排除する SPAM メールフィルタに利用されているナイーブベイズ分類器をはじめとする 6 種類の機械学習手法について，その性能を体系的に示すため，University of California, Irbine Machine Learning Repository より入手した英文コーパスと，独自に作成した日本語コーパスを用いて判別実験を行い，その結果を比較・考察した．英文コーパスの実験では各手法において訓練データ数を 500 ずつ増加させた際の判別率の推移を比較した．その結果，本実験においては Random Forest が最も判別性能が高い手法であることを確認した．また，SVM に関しては，8 種類のカーネル関数を用いてどの関数を用いるのが良いかを決めるため，同条件下で実験を行い，その結果を比較した．その結果，ANOVA カーネルが SPAM 判別に適したカーネルであることを確認した．日本語コーパス実験では，訓練データ数 1000 の時の判別性能を英文コーパスと条件を同期して判別実験を行い，その結果を比較した．その結果，作成した日本語コーパスにおける判別では SVM が最も SPAM 判別に適していることを確認した．

今後の展望として，英文コーパス実験ではスプラインカーネルにおける判別性能急落の原因を解明し，精度向上をはかる．また，日本語コーパス実験では，コーパスのデータ数増加，特徴量項目を再考し，コーパスの精度を上げて再実験を行う．この実験により，日本語という言語に特化した SPAM メールを判別するフィルタの作成，学習手法の考案が期待される．

謝辞

本研究を進めるにあたり，ご指導いただいた高知工科大学情報学群吉田真一講師に心から感謝致します．研究を進めるにあたって，まったく進捗のない私を見放さず，最後まで様々な観点からご指摘・ご指導いただきました．また，研究室活動においても，輪講における発表スライドの添削や各イベントの相談，飲み会でのお酒の飲み方など，様々なことを教えていただきました．深く感謝申し上げます．

本研究の副査を引き受けていただきました，高知工科大学情報学群島村和典教授と高知工科大学情報学群植田和憲講師に深く感謝いたします．島村教授には，発表直前に励ましのお言葉とお菓子をいただきました．発表や質疑に対する応答が非常に稚拙で不明瞭であったにもかかわらず，発表後に「良かったよ」のお言葉を頂いたときには，それまで再履修を覚悟して最低まで下がっていたモチベーションを取り戻すことができました．植田講師には，セッション終了後に稚拙な発表について謝罪に伺ったところ,「そんなことはない」とお言葉を頂きました．また，その後も発表した機械学習手法について 5 分ほど議論していただき，今後の研究に活かすことができました．島村教授と植田講師に深く感謝申し上げます．

同研究室の諸先輩方には，配属時の Free BSD のインストールからカスタマイズ，輪講の発表資料の指摘，飲み会でのお酒の飲み方など，様々なことを教えていただきました．深く感謝しております．

同期の 4 年生の皆さんには，研究の進捗具合，機械学習アルゴリズム構築についての助言を頂き，自分の研究を進めるにあたってモチベーションを保つことができました．また，研究以外に関しても，某 SNS ゲームで一丸となってプレーしたり，ギャンブルしに行ったりと，研究面以外でも非常に充実した生活を送ることができました．また，情報の研究室には稀な喫煙者が非常に多いメンバーで，一服に行くのに寂しさを感じない楽しいメンバーでした．私は進学するので残りますが，喫煙者が私を含め 2 人になってしまうのが寂しくてなりません．これを機に禁煙しようかとも思っています．

謝辞

同研究室の 3 年生の皆さんには，皆さんのあまりの優秀さに負い目を感じる面も多々ありましたが，研究について相談に乗っていただいたり，励ましていただいたりと，大変お世話になりました．今後も多い人で 2 年間，少ない人でもあと 1 年間研究室にいますが，変わらず接していただけたらと思います.

最後に，4 年間学費・生活費・精神面で支えてくれ，かつ更なる進学を許可してくれた家族に心から感謝いたします.

参考文献

[1] 株式会社シマンテック, ”シマンテック スパム&フィッシング マンスリーレポート 第 45 号”, 2010 年 9 月.

[2] Nello Cristianini, John Shawe-Taylor 著, 大北 剛 訳, ”サポートベクターマシン入門”,p.9, 共立出版株式会社.

[3] 金 明哲, ”R によるデータサイエンス”,p251, 森北出版株式会社.

付録 A

英文コーパスにおける判別実験結果のグラフ拡大図

ここでは，今までに示したグラフにおいて，多くの折れ線が重複していた部分についての拡大図を以下に示す.

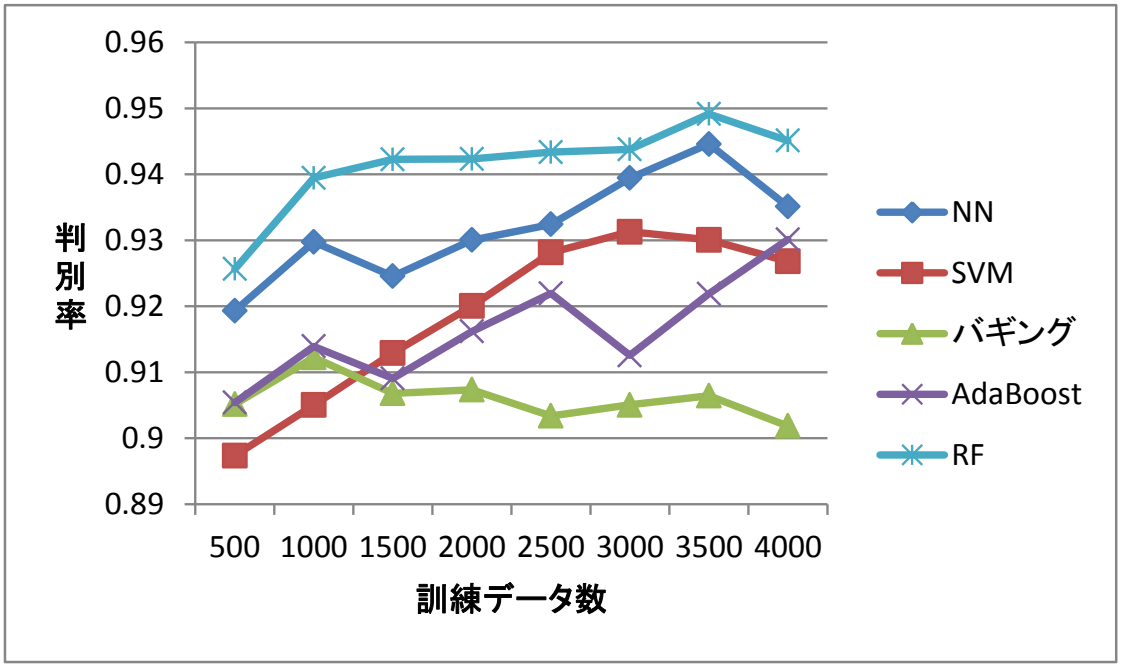

図 A.1　図 4.1 の拡大図

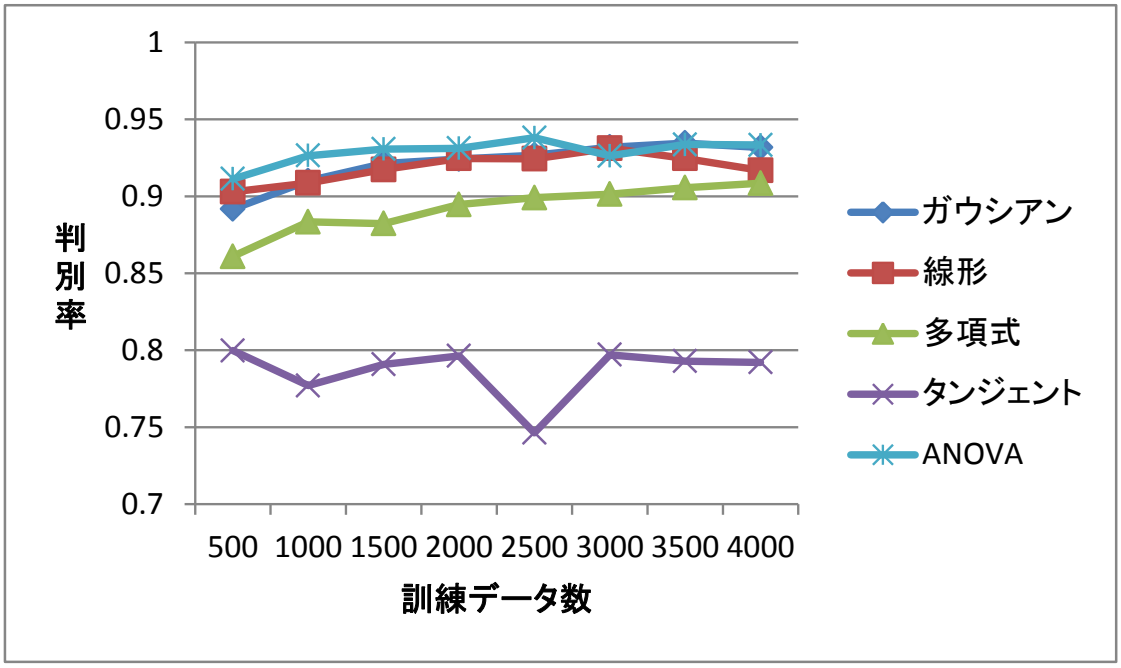

図 A.2　図 4.3 の拡大図

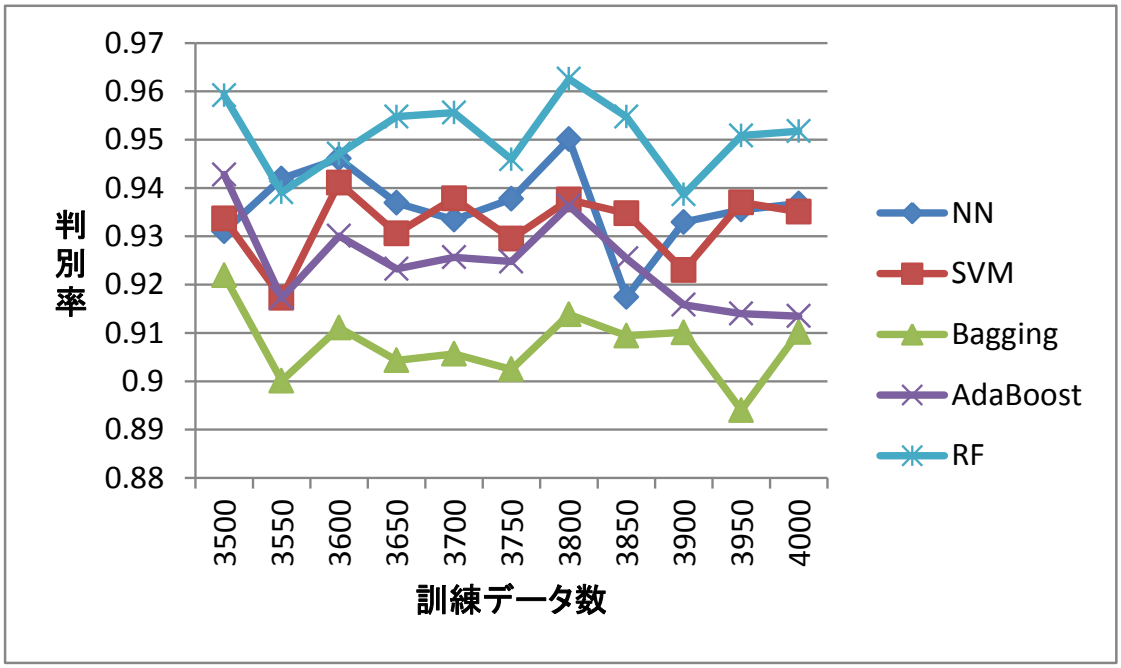

図 A.3　図 4.5 の拡大図