P2P ネットワークにおける Boosting を用いた信頼性の重み付け法について

2003MT019 久田 章太　　2003MT046 黒田 和樹
指導教員　河野 浩之

1　はじめに

P2P ネットワークはハイブリッド型とピュア型とに分類されるが，ピュア型 P2P ネットワークでは，悪意のあるピアが存在しやすいという問題点がある．そのため悪意のあるピアと取引しないようにするためには，そのピアが本当に信頼できるかどうかを取引する前に判断する必要がある．そこでピアの信頼性と評価が重要な要素となる．

この問題に関する研究として文献 [1] で荒井らは，Web 上であるキーワードで検索された集合を，ボトムアップクラスタリング，EM アルゴリズム，AdaBoost を用いた C4.5 により分類した．結論として，Boosting によるアンサンブル学習が分類において適しているのではないかという結論に至った．

また文献 [2] で Selcuk らは信頼ベクトルと信憑ベクトルを用い，未知のピアの信頼値を計算する方法を提案している．信憑ベクトルを使うことで未知のピアに対する信頼値の精度を向上させた．

本研究では文献 [1] と文献 [2] を用い，ピアの信頼性を判定する方法を提案する．特に未知のピアに対して評判を元に Boosting による信頼性判定を行う．

2　P2P 信頼性の関連研究

Selcuk ら [2] はユーザが扱うピアの信頼性を評価する P2P の評判に基づく信頼管理プロトコルの提案について紹介する．この提案はファイルをダウンロードする際，ピアの信頼値を用い，ダウンロード先を選択するものである．

ピアは直接通信をして結果を記録した信頼ベクトルと他ピアの評判が信頼できるかどうかを判断する信憑ベクトルを保持している．ダウンロード先を決める際に以下の手順をふむ．

1. 信頼ベクトルに登録されていれば，そこから信頼値を計算
2. 登録されていなければそのピアの評判を聞く
3. 評判と評判を返したピアの信憑ベクトルを用いて未知のピアの信頼値を計算
4. 信頼値の高いピアからファイルをダウンロード

本研究では，信頼ベクトル，信憑ベクトルを用いることで，ピアが信頼できるかどうか判断するための指標とする．

3　Boosting

Boosting とは精度の低い学習機 (弱学習機) を組み合わせることで，精度と汎化能力の高い学習機を構成するアンサンブル学習の手法である．Boosting の特徴は (a) 逐次的に学習機を構成，(b) 重み付きリサンプリング，(c) 弱学習機の重み付き結合である．代表的なものとして AdaBoost があげられる．

AdaBoost は教師データをある弱学習機に入力した際，その弱学習機が誤分類した教師データには高い重みを付け，正分類したデータには低い重みを付ける．この重みを指数関数的に付け，誤差が大きいデータにはより大きな重みを付けるのが AdaBoost の特徴である．この重み付き教師データをリサンプリングし，次の弱学習機の教師データとする．

また教師データの分類を終えた各弱学習機の誤り率を算出する．この誤り率からその弱学習機がどれくらい精度がよいかという信頼度を計算する．この手順を弱学習機の数だけ繰り返して，最終的に各弱学習機の重み付き多数決により結果を出す．AdaBoost アルゴリズムを図 1 に表す．

1. 重み $w_1(i) = \frac{1}{n}, (i = 1, 2, 3, ..., n)$ にする
2. For $t = 1, 2, ..., T$
   (1) 弱学習機 $f_t(x)$ を重み $w_t(i)$ を使って訓練データに適合させる
   (2) $error_t = \frac{\sum_{i=1}^{m} w_i(i) I(y_i \neq f_t(x))}{\sum_{i=1}^{m} w_t(i)}$ を計算する
   (3) $\alpha_t = \log \frac{1-error_t}{error_t}$ を計算する
   (4) 重み $w_{t+1}(i) = w_t(i) \exp[\alpha_t I(y_i \neq f_t(x_i))], (i = 1, 2, 3, ..., n)$ を更新する
3. $F(x) = sign(\sum_{t=1}^{T} \alpha_t f_t(x))$ を出力する

図 1: AdaBoost アルゴリズム

ここで n を教師データ数, 訓練用データを $(x_1, y_1) \cdots (x_i, y_i) \cdots (x_n, y_n)$ が与えられたとする．x_i は d 次元の入力ベクトル，y_i は x_i に対する出力ラベルであり，二値問題を扱うものとするので，$y_i = \{+1, -1\}$ とする．T は学習機の個数である．$sign$ 関数により，数値が正の場合は 1，負の場合は -1 を出力する．

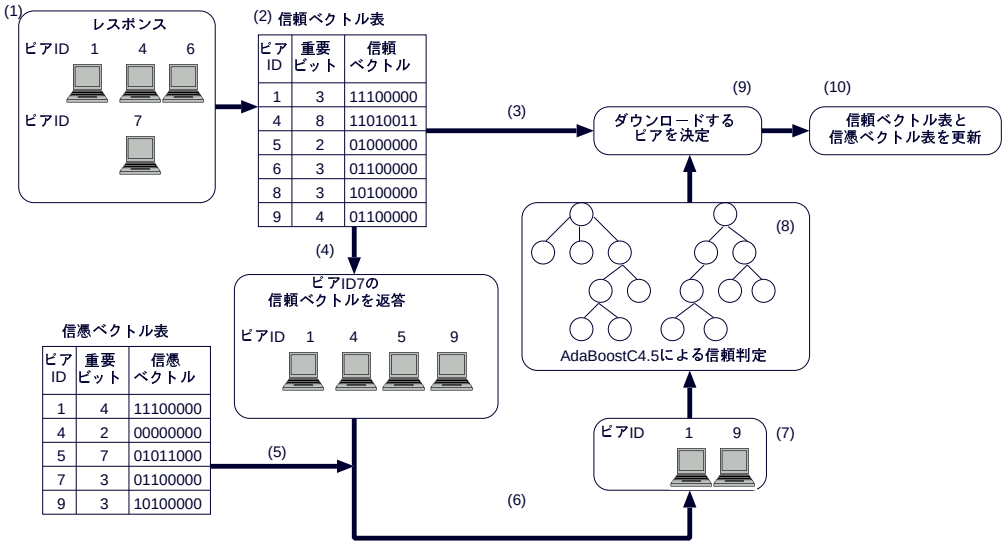

図 2: 提案した信頼性判別方法の概要

4 Boosting を用いたピアの信頼性判断の提案とアルゴリズム

未知のピアの信頼性判断を全て網羅しようとすると膨大なコストがかかってしまう．そこで本研究では，Selcuk らが提案した手法に AdaBoost を用いていくつかの教師データによる学習で効率的な学習を行う．

4.1 信頼ベクトルと信憑ベクトル

まず，本研究で使用する各ベクトルとその説明を以下に述べる．

信頼ベクトル　既知のピアの信頼性
信憑ベクトル　未知のピアの評判の信頼性
疑念ベクトル　既知のピアの信頼性の精度を向上
疑惑ベクトル　評判の信頼性の精度を向上

信頼ベクトル，信憑ベクトルはそれぞれ信頼ベクトル表，信憑ベクトル表に格納される．各ベクトル表にはピア ID，重要ビット，各ベクトルが格納されている．重要ビットとは信頼ベクトル，信憑ベクトルの上位何ビットを使用するかを表し，ピアとの通信を行うたびに更新していく．最大値は信頼，信憑ベクトルの長さである．また疑念ベクトルは信頼ベクトルを，疑惑ベクトルは信憑ベクトルを重要ビット分だけそれぞれ 0 と 1 を反転させることで生成する．

信頼ベクトルの更新はファイルをダウンロード後，重要ビットを 1 増やし，ファイルが目的のものであれば最上位ビットに 1 を，目的のものでなければ 0 を入れる．

信憑ベクトルの更新は以下の 2 つに分けられる．

(1) 未知のピアを信頼し，ファイルをダウンロードした場合
- 評判を用いたピアの重要ビットを 1 増やし，未知のピアが悪意のないピアならば信憑ベクトルの最上位ビットに 1 を挿入し，悪意あるピアならば 0 を挿入する．

(2) 未知のピアを信頼できないと判断した場合
- 評判を用いたピアの過去の振舞いが良ければ信憑値を少し増加させ，悪ければ少し減少させる．

また，信頼値，信憑値，疑念値，疑惑値を求めるには各ベクトルを以下の式 (1) でスカラー化する．

$$0 \leq \frac{(\text{各ベクトルの上位重要ビット分})_2}{2^{\text{重要ビット}}} < 1 \quad (1)$$

4.2 ピアの信頼性判定方法

本研究の提案を図 2 に示し，説明する．

(1) 自ピアがファイルクエリを送信後，複数のピアからレスポンスが来る．
(2) レスポンスが来たピアが信頼ベクトル表に登録されているか確認する．
(3) 登録されている場合は式 1 を元に疑念値計算，信頼値計算を行う．
(4) 登録されていない場合は他のピアに評判を聞く．
(5) レスポンスのあったピアの疑惑値，信憑値を計算する．
(6) 疑惑値により昇順にソートし，信憑値により降順にソートする．
(7) 上位ピアを抽出する．
(8) 抽出したピアから得た評判を元に Boosting にかける．
(9) (3) の結果と (7) の結果を元にダウンロードするピアを決定する．
(10) ダウンロード終了後，各ベクトル表を更新する．

4.3 ファイル交換分類アルゴリズム

実験プログラムのアルゴリズムを図 3 に示し，説明する．

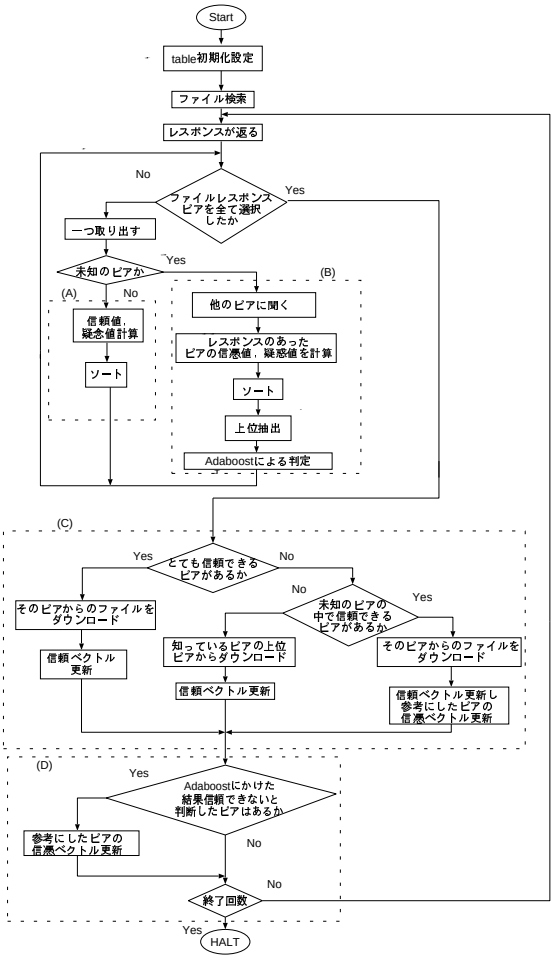

図 3: 実験プログラムのアルゴリズム

(A) ファイルレスポンスピアを知っている場合，信頼値と疑念値を求め，疑念値により昇順にソートし，信頼値により降順にソートする．

(B) ファイルレスポンスピアを知らない場合，信憑値と疑惑値を求め，疑惑値により昇順にソートし，信憑値により降順にソートする．その後上位ピアを抽出し，Boosting により “信頼できる” か “信頼できない” かを判断する．

(C) ファイルをダウンロードするピアを決定し，ベクトルを更新する．

(D) 未知のピアで信頼できないと判断するのに用いた評判を返したピアの信憑ベクトルを更新する．

5 Boosting を用いたピアの信頼性判断の実験

5.1 C4.5 決定木，SVM による分類

“信頼できる” 540 個，“信頼できない” 60 個の計 600 個の教師データを用い，C4.5 決定木と SVM 単独で分類器を作成した．各分類器の分類率の結果を表 1 に示す．

表 1: 弱学習機単体による分類結果

	C4.5	SVM
トレーニングデータ	90%	94%
Cross Validation	90%	89.6667%

C4.5 決定木は 90%の分類率が出ているが，深さ 1 で全て “信頼できる” に分類する決定木を作成していた．これは使用した教師データの 90%が “信頼できる” であったため，決定木を作成しなくても高い分類結果が得られると判断したのだと思われる．

SVM では 89.6667%の分類結果であった．しかしこの分類結果ならば，全て “信頼できる” と分類した方が高い分類結果が得られるため，SVM による分類は効果的ではない．教師データに偏りがある場合では，分類器単独で教師データを分類することは難しいことが分かった．

5.2 AdaBoost を用いた場合の性能評価

分類器単独での分類が困難であったので，上記の C4.5 決定木と SVM を弱学習機とした AdaBoost を用いて分類器を作成した．結果を表 2 に示す．

表 2: Boosting による分類結果

	AdaBoostC4.5	AdaBoostSVM
トレーニングデータ	100%	97%
Cross Validation	90.1667%	86.5%

AdaBoostC4.5 の場合，Cross Validation による分類率が 90%を越えたため，トレーニングデータに用いなかったテストデータでもある程度対応可能できることが分かる．

AdaBoost により作成された 10 個の木のなかに，C4.5 決定木単独の時に作られた深さ 1 で全て “信頼できる” に分類する木も作成されていた．しかしその木に対する重みは 2.2 で，他の木に対する重みが 3 以上であることを考慮してもとりわけ低かった．AdaBoostC4.5 は偏った教師データを用いてもよい分類器を作成できることが分かった．

一方 AdaBoostSVM の場合，Cross Validation による分類が SVM 単独の場合より 3.1667%低くなり，Boosting を用いることで逆に性能が落ちてしまった．その理由として AdaBoost の特性である “うまく分類できなかったデータに対して指数関数的に重みを付ける” という重み付け方が適していなかったのではないかと思わ

れる．

また，教師データの “信頼できる”，“信頼できない” の比率を 9:1 から 99:1 に変更したときの分類結果を表 3 に示す．AdaBoostC4.5，AdaBoostSVM ともに Cross

表 3: 教師データをより偏らせた時の分類

	AdaBoostC4.5	AdaBoostSVM
トレーニングデータ	100%	100%
Cross Validation	98.8333%	98.8333%

Varidation を用いた分類が 98.8333%であった．これは教師データに含まれる “信頼できない” の個数が 6 個ととても少なく，Cross Validation により教師データを分割した際に全部 “信頼できる” となったデータ群が存在したためだと思われる．以上のことから AdaBoostC4.5 を使うことで未知のピアに対して効果のある結果を得られることがが分かった．

5.3 先行研究 [2] との比較

5.3.1 naive 環境

悪意のあるピアの比率を 1%と 10%と変えた naive 環境による比較を行った．シミュレーション結果を図 4 に示す．提案した方法では悪意ピアからのダウンロード数

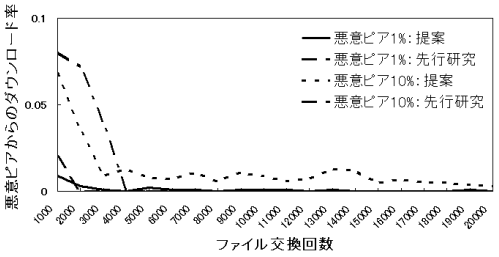

図 4: naive 環境による比較

は徐々に減ってはいくが 0 にはならなかった．これは先行研究が評判の数値計算によりピアの信頼性を確保しているのに対し，提案した手法では評判の組み合わせにより信頼性を確保しているためであると考えられる．

しかし未知のピアとの取引が多いファイル交換回数の少ない初期の段階においては，我々が提案した手法の方が悪意のあるピアからのダウンロードが少ない．このことより提案した手法は初期の段階から悪意のあるピアとの取引を制限することが出来たといえる．

5.3.2 hypocritical 環境

次に hypocritical 環境による比較を行った．悪意ファイルを送る確率を 10%，25%とした場合の悪意のあるピアの比率を 1%，10%と変えてシミュレーションを行った結果を図 5 と図 6 に示す． 提案した手法は初期の段階から悪意のあるピアからのダウンロード回数を低く抑えることができた．しかし悪意のあるピア 10%の環境では，ファイル交換を重ねても 0 になることはなかった．

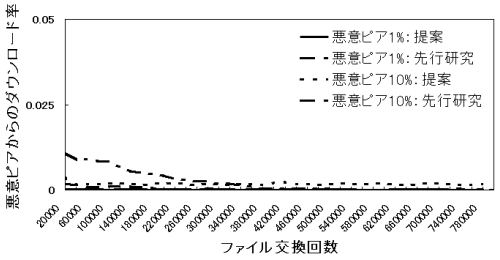

図 5: 悪意ファイルを送る確率 10%

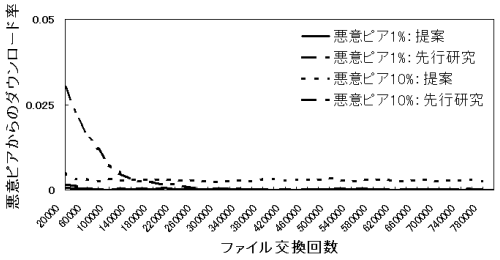

図 6: 悪意ファイルを送る確率 25%

また，悪意のあるピアの存在比率が同じ場合，悪意ファイルを送る確率に関係なく同じようなグラフになった．

このことから本システムは悪意のあるピアの性質には大きな影響を受けず，また悪意のあるピアの存在率により多少悪意ピアからのダウンロードが上下するものの大きな変化がないため，ネットワーク環境に大きく左右されないロバストな手法であることが分かった．

6 まとめ

我々は信頼ベクトルや信憑ベクトルを用いた Boosting によるピアの信頼性判定方法を提案した．その結果，AdaBoostC4.5 は教師データに大きな偏りがあっても分類可能であり，初期の段階から悪意のあるピアとの取引を抑制することができた．そのため，大多数のピアが悪意の無いピアである P2P ネットワークにおいて，Boosting による信頼性評価は非常に効果的であるといえる．

参考文献

[1] 荒井幸代，村上陽平，杉本悠樹，田仲正弘，“Boosting を用いた評判の信頼性評価方法，” 人工知能学会研究会資料，第 4 回セマンティックウェブとオントロジー研究会資料集，SIG-SW&ONT-A302，2003．

[2] A．A．Selcuk，E．Uzum，M．R．Pariente，“A Reputation-Based Trust Management System for P2P Networks，” IEEE International Symposium，pp.251-258，19-22 April 2004．