RF-008

単一拡散系列からの期待影響度曲線の推定
Estimating Expected Influence Curve from Single Diffusion Sequence

吉川 友也† 斉藤 和巳† 元田 浩‡ 木村 昌弘§ 大原 剛三※
Yuya Yoshikawa Kazumi Saito Hiroshi Motoda Masahiro Kimura Kouzou Ohara

1. はじめに

インターネットや World Wide Web の興隆は，大規模な社会ネットワークの発生を加速している．したがって，情報を普及させるための重要メディアとして，最近，社会ネットワークが注目されている[1].

情報拡散の基本確率モデルは IC モデル[2]である．私たちは現実世界の情報拡散により近くなるように連続時間に対応した CTIC モデル[3]によって情報拡散分析を行っている．観測された拡散系列が CTIC モデル上で情報拡散したと仮定すると，拡散確率 κ と時間遅れパラメータ r を学習することができる．しかし，詳しくは後述するが，私たちが行なった拡散シミュレーションによると，観測された拡散系列は同じパラメータを使っているにもかかわらず，ばらつきのある結果が得られている．これは現実の場合にも起こりえる現象で，加えて，ほとんどの場合単一の拡散系列しか観測できない．このような条件下で，いかに正確に期待影響度曲線を計算するかが私たちの課題である．

そこで本論文では，単一の拡散系列より期待影響度曲線を推定する方法を提案する．詳細には，得られた拡散系列より CTIC モデルのパラメータを学習し，その学習パラメータによって期待影響度曲線を推定する．また，評価実験として，シミュレーションによって作った拡散系列群を用いて，頑健に期待影響度曲線を求めることができるかを検証する．評価実験により，期待影響度曲線を頑健に推定できることが保証されれば，現実の拡散系列を用いた情報拡散分析において，今回の推定法で確からしい期待影響度曲線を求めることが可能であることが示唆される．

2. 情報拡散モデルの CTIC モデル

まず，使用する情報拡散モデル CTIC モデルを定義する．有向ネットワークを $G=(V,E)$ で定義する．$V=\{u, v, w, \cdots\}$ はノード集合を，$E=\{(u,v),(v,w),\cdots\}$ はリンク集合を表す．また，ノード v がリンクする子ノード集合を $F(v)=\{w; (v,w)\in E\}$ とする．ここで，ノードが情報を保持している状態をアクティブと呼び，そうでない状態を非アクティブと呼ぶ．CTIC モデルでは，非アクティブからアクティブへ状態は変わるが，逆は起こらない．アクティブなノード v は，各出リンクを通し独立に子ノード集合 $F(v)$ の各ノードを確率 κ $(0\leq\kappa\leq1)$ でアクティブにすることができる．この情報拡散試行が行われるのは一度限りで，時刻 t_v で子ノード w をアクティブにする試行に成功したとき，w がアクティブになる時刻は指数分布 $p(t)=r\exp(-r(t-t_v))$ で与えられるとする．r は時間遅れパラメータを表す．なお，リンク毎に，拡散確率や時間遅れパラメータが異なるように一般化したモデルも同様に定義できる．

† 静岡県立大学 University of Shizuoka
‡ 大阪大学 Osaka University
§ 龍谷大学 Ryukoku University
※青山学院大学 Aoyama Gakuin University

3. 期待影響度曲線の推定法

今回提案する期待影響度曲線の推定法について説明する．この推定法は入力値として観測した拡散系列 d，情報源ノード v_0 をとり，出力値は期待影響度曲線 $c(t;v_0,d)$ となる．以下は推定法のアルゴリズムである．

1. 拡散系列 d よりモデルパラメータ κ, r を学習する
2. κ, r より，ノード v_0 を情報源として M 回拡散シミュレーションを行い，人工拡散系列群 $S=\{s_1, s_2, \ldots, s_M\}$ を求める
3. S の平均として期待影響度曲線 $c(t;v_0,d)$ を求める

人工拡散系列 $s_m\in S$ は，時刻 t にアクティブになったノード v の対 (t, v) で構成され，以下で表現する。

$$s_m=\{(v_0,t_0),(v_{m,1},t_{m,1}),\ldots,(v_{m,T},t_{m,T})\},m=1,\ldots,M.$$

このとき時刻 t における S の平均 $c(t;v_0,d)$ は，

$$c(t;v_0,d)=\frac{1}{M}\sum_{m=1}^{M}|\{(v,\tau)\in s_m;\tau\leq t\}| \quad (1)$$

となる．なお，パラメータ学習法は先行研究[3]に従う．

4. 評価実験

以下の評価法を用いて実験を行なう．また，ネットワークの可視化により拡散系列 d の拡散現象について評価する．

4.1 評価法

今回の提案法の有用性を評価するために，人工データを用いた実験を行なう．評価は以下の方法で行なう．

1. 真の κ^*，r^* を決め，N 回のシミュレーションを行い，拡散系列群 $D=\{d_1, d_2, \ldots, d_N\}$ を生成する
2. 各 $d_n\in D$ の影響曲線 $\varphi(t;v_0,d_n)$ を以下で計算する
$$\varphi(t;v_0,d_n)=|\{(v,\tau)\in d_n;\tau\leq t\}|$$
3. D の平均値である c^* を以下より求める．
$$c^*(t;v_0)=\frac{1}{N}\sum_{n=1}^{N}|\{(v,\tau)\in d_n;\tau\leq t\}| \quad (2)$$
4. 各 n=1, …, N において，拡散系列 d_n から κ_n, r_n を学習する
5. 各 n=1, …, N において，M 回拡散シミュレーションを行い，人工拡散系列群 $S_n=\{s_{n,1}, s_{n,2}, \ldots, s_{n,M}\}$ を求める
6. 各 n=1, …, N において，S_n の平均値 c_n を式(1)で求め，期待影響度曲線群 $C=\{c_1, c_2, \ldots, c_N\}$ とする
7. c^* からの誤差である E_D と E_C を以下の式で求める．

$$E_C(t)=\sqrt{\frac{1}{N}\sum_{n=1}^{N}(c(t;v_0,d_n)-c^*(t;v_0))^2}$$

$$E_D(t)=\sqrt{\frac{1}{N}\sum_{n=1}^{N}(\varphi(t;v_0,d_n)-c^*(t;v_0))^2}$$

最終的には誤差曲線 E_C, E_D によって評価を行なう．

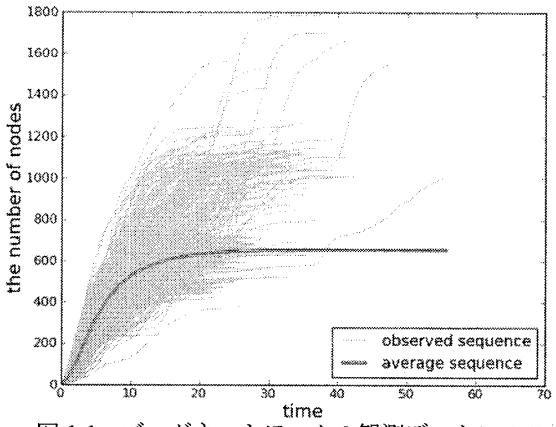

図 1-1　ブログネットワークの観測データ(κ=0.1)

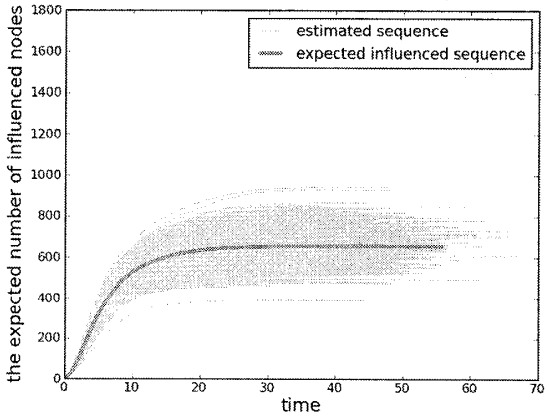

図 1-2　ブログネットワークの期待影響度

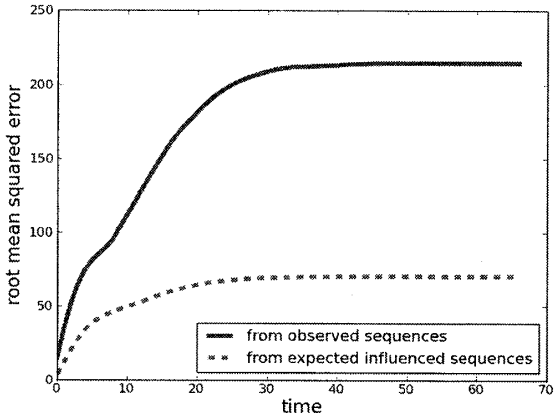

図 1-3　ブログネットワークの真の影響度からの誤差

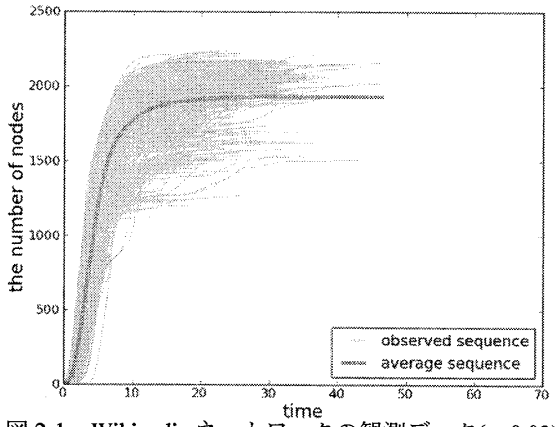

図 2-1　Wikipedia ネットワークの観測データ(κ=0.03)

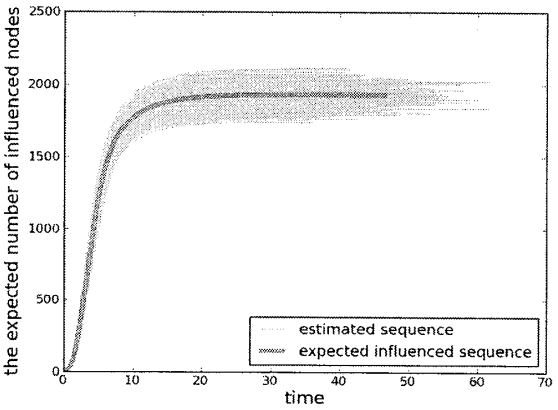

図 2-2　Wikipedia ネットワークの期待影響度

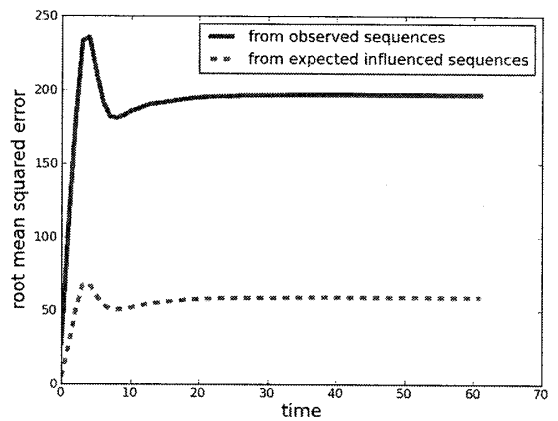

図 2-3　Wikipedia ネットワークの真の影響度からの誤差

4.2 実験データ

実験データとして，ブログと Wikipedia のネットワークを使用する[5]．これらのネットワークは，社会ネットワークの特徴であるスケールフリー性[4]を持ち，実験データとして妥当だと考えられる．ブログネットワークはノードを記事，リンクをトラックバック関係とした繋いだもので，ノード数は 12,047，リンク数は 79,920．Wikipedia ネットワークは「人名一覧」からの人物ネットワークで，ノード数は 9,481，リンク数は 245,044 である．

4.3 実験設定

実験を行なう前に，CTIC モデルの挙動を決める真のパラメータを決定する．今回の実験では，ブログネットワークは拡散確率 κ^*=0.1，時間遅れパラメータ r^*=1.0 とし，Wikipedia ネットワークは拡散確率 κ^*=0.03，時間遅れパラメータ r^*=1.0 とする．また，情報源となるノードを 1 つ決定する．今回の実験では，各ネットワークでの最終的な期待影響度の最も高いノードを情報源とする．また，拡散系列群 D の個数 N=1,000，期待影響度曲線を求めるためのシミュレーション回数 M=100 とする．

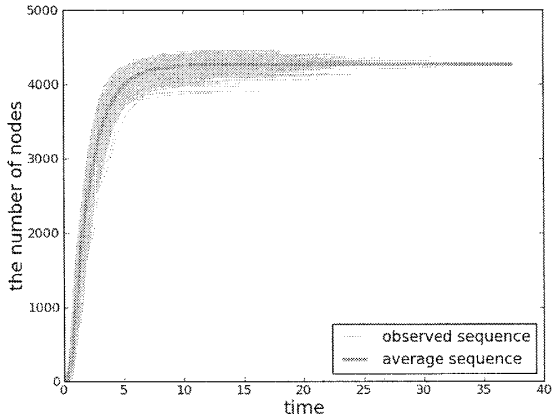

図 3-1　Wikipedia ネットワークの観測データ(κ=0.09)

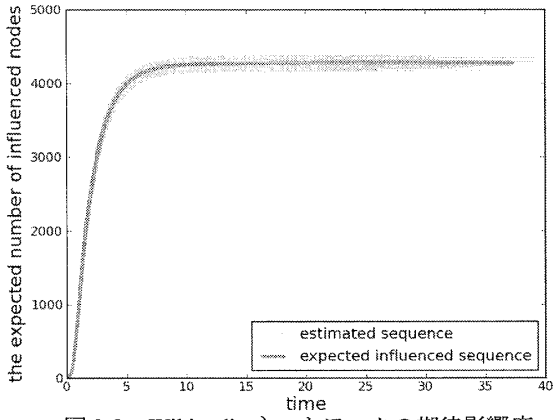

図 3-2　Wikipedia ネットワークの期待影響度

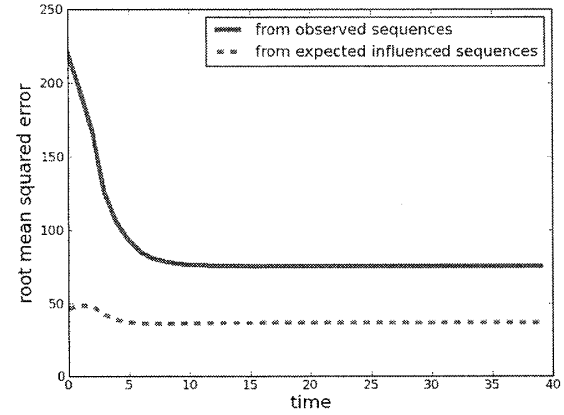

図 3-3　Wikipedia ネットワークの真の影響度からの誤差

4.4 実験結果

図 1-1 から図 1-3 までがブログネットワークに関する一連の実験結果を表し，図 2-1 から図 2-3 までが Wikipedia ネットワークに関する一連の実験結果を表す．まず，図 1-1 と図 2-1 は各ネットワークで拡散シミュレーションをした結果得られた観測系列群のアクティブノード数の分布図である．横軸は時刻を表し，縦軸は時刻 t までにアクティブになったノードの総数を表す．1 本の細線は 1 つの観測系列であり，1,000 本の細線が描かれている．この図を見ると，図 1-1 では，最終的に 300 程度のノードしかアクティブにならないときもあれば，2,000 以上のノードがアクティブになるときもあり，同じパラメータからでも幅のあるデータが得られることが分かる．これと同様の結果が図 2-1 でも見られる．そして，太線は細線の平均値 c^*であり，これを真の影響度と定義する．次に，各拡散系列より情報拡散の仕方を決めるパラメータである，拡散確率と時間遅れパラメータを学習する．私たちの実験の結果，この学習パラメータは真のパラメータの周辺に分布することを確認している[6].

また，図 1-2 と図 2-2 は，学習パラメータを用いて拡散シミュレーションを行った結果である．細線は図 1-1，図 2-1 同様にアクティブノードの分布を表すが，この図では 1 本の線が期待影響度曲線群である．また，太線は真の影響度曲線 c^*を表す．図 1-2 のブログネットワークの場合，最終的な期待影響度の最大値は 800 程度，最小値は 400 程度であり，最初に得た拡散系列群に比べると，真の影響度曲線 c^*に対しての信頼性の高さがうかがえる．これは図 2-2 の Wikipedia ネットワークにおいても，同様の結果が見られる．

最後に，拡散系列群 D と期待影響度曲線群 C の真の影響度曲線 c^*からの差を検証し，定量的に期待影響度曲線の信頼性を検証する．図 1-3 と図 2-3 が各ネットワークでの結果を示し，横軸は時刻，縦軸は各時刻における真の影響度曲線との RMSE（Root Mean Squared Error）を表す．また，実線は拡散系列群からの差 E_D，破線は期待影響度曲線群からの差 E_C を示す．この実験結果より，拡散系列群よりも期待影響度曲線群の方が，真の影響度曲線 c^*に対しての誤差が小さいことが明らかである．

4.5 真のパラメータを変更した場合の実験結果

変更後の真のパラメータは，ブログネットワークは拡散確率 κ^*=0.3，時間遅れパラメータ r^*=1.0，Wikipedia ネットワークは拡散確率 κ^*=0.09，時間遅れパラメータ r^*=1.0 と設定する．時間遅れパラメータを変更しなかったのは，私たちは実験より，時間遅れパラメータはネットワーク上での拡散スピードを決定するだけで，最終的な影響度にはほとんど影響がないということを確認しているからである．今回の研究では最終的な影響度に着目しているので，時間遅れパラメータは変更せずに，直接的に影響度を決める拡散確率のみを変更した．

図 3-1 から図 3-3 は Wikipedia ネットワークでの実験結果を示す．ブログネットワークはページ数の制限により割愛するが，Wikipedia ネットワークと同様の結果が得られている．図 3-1 は，拡散シミュレーションをした結果得られた拡散系列群 D のアクティブノード数の分布図である．この図を見ると，各細線が真の影響度曲線を表す太線にまとわり付くように分布していることが分かる．すなわち，1 つ 1 つの拡散系列がほぼ同じような結果であり，これは同じネットワークの結果である図 2-1 に比べると大きな違いがある．図 3-2 は学習パラメータを用いて拡散シミュレーションを行った結果である．1 本の細線は 100 回の拡散シミュレーションの平均値で期待影響度曲線を表す．図 3-3 は拡散系列群と期待影響度曲線群の真の影響度曲線との差を表す．拡散確率を変える前の実験結果である図 2-3 と同様に，期待影響度曲線群の方が拡散系列群よりも真の影響度曲線からの誤差が小さいが，図 2-3 よりは両者の差は小さい．

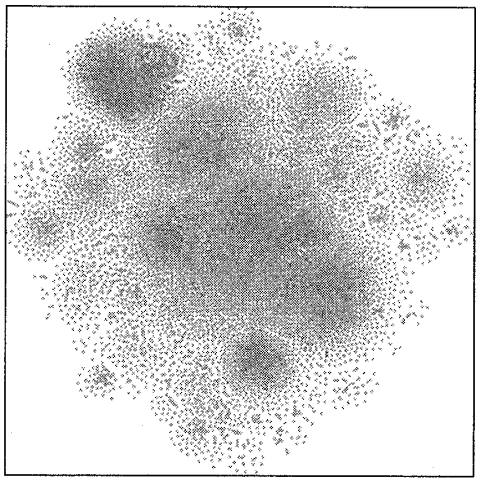

図 4-1　情報がよく拡散する場合

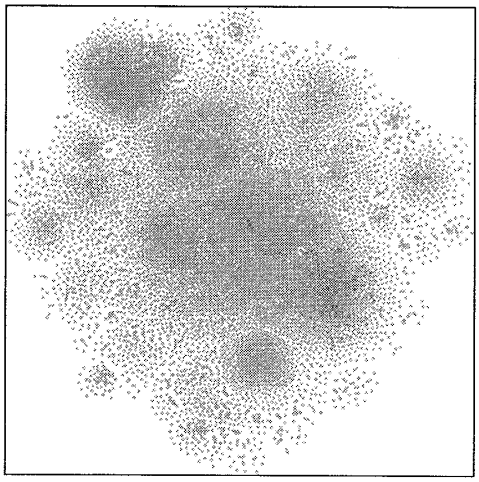

図 4-2　情報が拡散しない場合

4.6 ネットワーク可視化による評価

最後に，各ネットワークをクロスエントロピー法[7]によって可視化する．図 4-1 はブログネットワークで最終的にアクティブノード数が最も多かった試行での可視化結果，図 4-2 はブログネットワークで最終的にアクティブノード数が最も少なかった試行での可視化結果である．また，ノードを表す点が密集する部分はノード間の連結が強いことを表し，ここにはコミュニティが存在すると考えることができる．両者を比較すると，図 4-2 では情報源ノードの属するコミュニティではアクティブノードが存在するものの，他のコミュニティへの情報拡散はまばらである．対して，図 4-1 では情報源ノードの属するコミュニティでアクティブノードが多く存在し，かつ，他のコミュニティにもアクティブノードが散らばっていることがわかる．この結果，図 4-1 の試行では最終的なアクティブノード数が 1,789 だったのに対し，図 4-2 の試行ではアクティブノード数は 220 にとどまる結果となっている．また，Wikipedia ネットワークにおいても同様の結果が得られている．

5. 考察

今回の実験では，ブログネットワークと Wikipedia ネットワークで拡散シミュレーションを行い，これを拡散系列群とした．これによって得られた 1 つ 1 つの拡散系列によれば，図 1-1 や図 2-1 のように，アクティブノード数にはばらつきがあることがわかる．私たちは，このような結果になる原因はネットワークのコミュニティ構造にあると考える．コミュニティ構造について簡単に定義すると，コミュニティ内ではリンクは密に存在し，コミュニティ間ではリンクは疎であるような構造を表す．これはコミュニティ内では情報は伝わりやすいが，コミュニティ間では情報が伝わりにくいことを意味する．今回の実験の場合，アクティブノード数がばらつく理由はコミュニティ間を情報が伝わったからであり，伝わるときにはそのコミュニティ内で情報はより多くのノードに伝わり，伝わらないときには，情報は少ないノードにしか伝わらないことになる．これは実験結果である図 4-1，図 4-2 を比較しても明らかである．また，真のパラメータを変更して実験を行なうと，図 3-1 で示したように拡散系列群のばらつきがなくなり，何度シミュレーションを行っても最終的なアクティブノード数がある値に近くなっていることが分かる．これは，拡散確率を上げることによりコミュニティ間を情報が伝わりやすくなり，拡散確率の低いときに起きていたコミュニティ間を情報が渡ったり渡らなかったりといったような，確率的な揺らぎが起こりにくくなっているためと考えられる．

6. おわりに

本論文では，ウェブなどの社会ネットワークから得られた拡散系列より，期待影響度曲線を推定する方法を提案した．この提案法はパラメータの学習結果を期待影響度曲線で表した点で新しく，拡散能力をアクティブノード数という比較可能な形で表現することが可能になった．また，この提案法の評価実験として，現実のネットワークを使った真の影響度曲線と期待影響度曲線群の誤差計算をシミュレーションによって行った．その結果，1 回 1 回で得られる拡散系列の結果がばらつく場合には，期待影響度曲線群は拡散系列群と比べ，真の影響度曲線からの誤差が小さい．また，拡散系列群の結果がばらつかない場合には，ばらつく場合ほどではないが，それでも確かに期待影響度曲線群の方が誤差は小さい．したがって，評価実験より，私たちの提案法は確からしい期待影響度を求めることが可能だと示唆されたと考える．

参考文献

[1] Leskovec, J., Adamic, L., Huberman,B. A.,"The dynamics of viral marketing", EC'06, 228–237 (2006).
[2] Kempe, D., Kleinberg, J., Tardos, E., "Maximizing the spread of influence through a social network.", Proceedings of the 9th ACMSIGKDD International Conference on Knowledge Discovery and Data Mining(KDD-2003)., 137-146 (2003).
[3] 吉川 友也, 伏見 卓恭, 斉藤 和巳, 元田 浩, 木村 昌弘, "社会ネットワーク上での情報拡散データの分析", 第 8 回情報科学技術フォーラム (2009).
[4] Barabasi, A., Albert, R., "Emergence of Scaling in Random Networks", Science, Vol.286, No.5439, pp.509-512 (1999).
[5] Kimura, M., Saito, K., and Motoda, H., "Blocking links to minimize contamination spread in asocial network", ACM Transactions on Knowledge Discovery from Data, Vol. 3, No. 2, Article 9 (2009)
[6] 吉川 友也, 斉藤 和巳, 元田 浩, 木村 昌弘, 大原 剛三, "拡散データからのモデル推定による期待影響度の予測", 情報処理学会創立 50 周年記念（第 72 回）全国大会 (2010).
[7] Yamada, T., Saito, K., Ueda, N., "Cross-entropy directed embedding of network data", Proceedings of the 20th International Conference on Machine Learning, pp.832-839 (2003).