情報利得を用いたラベル選定に基づくマルチクラスタリング手法の評価

加藤　大智†　　渡辺　陽介††　　横田　治夫†

† 東京工業大学大学院 情報理工研究科 計算工学専攻
†† 東京工業大学 学術国際情報センター
〒 152-8550 東京都目黒区大岡山 2-12-1
E-mail: †kato@de.cs.titech.ac.jp, ††watanabe@de.cs.titech.ac.jp, †yokota@cs.titech.ac.jp

あらまし　インターネットの普及などによるデータ量の増加から，大量のファイルを高精度で分類することが求められている．しかし教師なしのクラスタリングでは利用者の意図しない分類になることが多い．そこでファイル毎に付与されたキーワードを利用することを考える．本研究では，クラスタリング精度向上のため，キーワードの中から分類のラベルとして適するものを情報利得を指標として選定する．この結果を基にキーワードがないファイルを SVM に基づく半教師付きクラスタリング手法を用いて分類する．2 値クラスタリングの結果をマルチクラスタリングに適用するにはいくつかの戦略が考えられるため，本研究では論文ファイルの分類を通してこれらを比較し評価する．

キーワード　クラスタリング, クラスタラベリング, SVM, 情報利得, リサーチマイニング

1. はじめに

近年，インターネットの普及などにより，さまざまなところで情報量が爆発的に増加している．しかし，このような大量の情報を人力で分類するのはコストが非常に高い．そこで，大規模データから傾向や新たな知見を得るための手法として，似たデータをグループ分けするクラスタリングと呼ばれる手法がよく用いられている．代表的なものとしては，k-means 法や凝集型階層的クラスタリングがある．

クラスタリングの特徴の一つとして，教師データをあらかじめ与えない，教師無しの手法であることがあげられる．このため，事前にどのような分類があるか分からない状態からデータの傾向を見るようなときに用いられるが，データの類似性のみを用いて判定を行うため，利用者の意図とは異なる分類となることも多い．特に，画像や論文などのファイルをテーマごとにクラスタリングする場合では，文字認識などのパターン認識とは違い類似性の定義も曖昧であり，最適な特徴量を選択するのが難しいことから，ファイルの内容のみから直接クラスタを生成するのは困難である．これを解消するため，クラスタリングの際にラベル付きデータ（どのクラスに属するか判明しているデータ）を部分的に与えてクラスタリングを行う手法があり，これは半教師付きクラスタリングと呼ばれている．しかしファイルのクラスタリングでは，利用者が望む分類に対してラベル付きのファイルをあらかじめ与えることは難しい．

そこで，厳密な教師データを人手で与える以外のアプローチとして，本研究ではファイルに付けられたキーワードやタグを分類指標として用いることを考える．キーワードが付けられているファイルとしては，動画や画像，書籍，論文などがあげられる．ただしキーワードはすべてのファイルについているとは限らず，またキーワードの中には分類に適さないものもあるため，単純にキーワードのみを用いてすべてのファイルを分類することは難しい．

以前我々は，この問題を解決するため，キーワードの情報を選別し有用なものだけを残して擬似教師データを作り，これと半教師付きクラスタリングを組み合わせて精度の高いクラスタリングを行うことを提案した [1]．また，精度の高いクラスタリングを行うため，擬似教師データを制約としてサポートベクターマシン (SVM) で分離超平面を求め，その結果を初期値としてマージン最大化クラスタリングを行う手法と，擬似教師データを用いて半教師付きの SVM を実行する手法を提案した [2]．しかしこれらは 2 値クラスタリングの手法であり，この手法をマルチクラスタリングに拡張する必要があった．本研究ではマルチクラスタリングを実現するため，2 値クラスタリングの手法を多段に使用し階層的にファイル集合を分割する手法と，2 値クラスタリングの結果を用いて連続的な決定関数を求め非階層的に分割する手法を用いることを提案する．また論文ファイルの分類を通してこれらを比較し評価する．

本論文の概要は以下の通りである．まず第 2 章で本研究で使用する SVM を用いたクラスタリング手法について述べる．第 3 章で評価対象とするクラスタリング手法を説明する．第 4 章で評価実験の手法と実験結果の提示，考察を行う．第 5 章で関連研究を紹介し，第 6 章でまとめと今後の課題を述べる．

2. 準　　備

まず，本研究のクラスタリングで用いるサポートベクターマシン [3]，半教師付きサポートベクターマシン [4]，マージン最大化クラスタリング [5] の概要を説明する．

2.1 SVM

サポートベクターマシン (Support Vector Machine, 以下 SVM) [3] とは，正例と負例の 2 クラスからなるラベルが付い

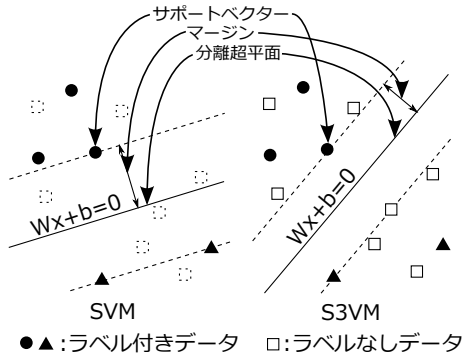

図 1: SVM,S3VM における分離超平面の求め方

た教師データから，その 2 つのクラスを線形分離する超平面を求める手法である．図 1 の左では，正例を丸，負例を三角で表した教師データを分離するような超平面を求めている．点線の四角で表した，ラベルがないインスタンスは分離超平面の決定に用いることができない．超平面の式を $D(\mathbf{x}) = \mathbf{W}^T\mathbf{x} + b = 0$ とすると，$D(\mathbf{x}_i) > 0$ となるインスタンス $\mathbf{x}_i$ のラベルは $+1$，$D(\mathbf{x}_i) < 0$ となる $\mathbf{x}_i$ のラベルは -1 となる．このような，インスタンスが属するクラスを決定させる関数 $D(\mathbf{x})$ を決定関数と呼ぶ．インスタンスを正しく分離できるような超平面はいくらでも考えられるが，SVM では最も近いサンプルとの距離（マージン）が最大となるような超平面を解とする．しかし教師データを線形分離できる超平面が存在しない場合もある．本研究では，多少の識別の誤りを許容するソフトマージンと呼ばれる手法を用いた．

SVM では，教師データから求めた分離超平面を用いて，クラスが未知のインスタンスを分離する．ただし，その際の未知インスタンス集合の分布に合わせて分離超平面を微調整することはできない．

2.2 S3VM

SVM を半教師付きクラスタリングに拡張する手法として，半教師付きサポートベクターマシン (Semi-Supervised Support Vector Machine, 以下 S3VM) と呼ばれるアルゴリズムが提案されている [4]．S3VM では，ラベルがついていないインスタンスが何らかのクラス分類を持つと考え，とり得るすべてのクラス分類について分離超平面を求め，マージンが最大のものを，集合を分離する超平面とする（図 1 の右）．

ラベル付きデータを $x_1, \ldots, x_l$，ラベルなしデータを $x_{l+1}, \ldots, x_n$，ラベル付きデータのラベルを $y_i \in \{-1, +1\}$[1] とすると，S3VM は，次のようにマージンの逆数 $\mathbf{W}$ を最小化する問題として定式化することができる．

$$\min_{y_{l+1},\ldots,y_n} \min_{\mathbf{W},b,\xi_i} \frac{1}{2}\mathbf{W}^T\mathbf{W} + \frac{C_l}{n}\sum_{i=1}^{l}\xi_i + \frac{C_u}{n}\sum_{j=l+1}^{n}\xi_j \quad (1)$$

$$s.t.\ y_i(\mathbf{W}^T\mathbf{x}_i + b) \geqq 1 - \xi_i, \forall i = 1, \ldots, l$$

$$y_i(\mathbf{W}^T\mathbf{x}_j + b) \geqq 1 - \xi_j, \forall j = l+1, \ldots, n$$

$$\xi_i \geqq 0, \forall i = 1, \ldots, l$$

$$\xi_j \geqq 0, \forall j = l+1, \ldots, n$$

ここで，C_l はラベル付きデータの，C_u はラベルなしデータの識別誤りの許容度を調整するパラメータである．C_l, C_u が大きいほど，識別の誤りを許容しなくなる．

S3VM の特徴は，ラベルつきデータのみから求められる超平面をそのまま用いるのではなく，ラベルが未知のインスタンスの分布に合わせて超平面を設定できることである．図 1 で言えば，SVM では分離超平面がラベルなしデータの近くに引かれてしまっているが，S3VM ではラベルなしデータも考慮した位置に分離超平面を引くことができている．

本研究では，これを高速に解く手法である CutS3VM [4] を用いてクラスタリングを行う．CutS3VM では，ランダムに初期平面を与えその周辺のインスタンスのみを用いて新たな分離超平面を求めることで，計算量の削減を図っている．

2.3 MMC

教師データが全く存在しない集合に対しては，SVM の考え方を教師なし 2 値クラスタリングの手法として利用したマージン最大化クラスタリング (Maximum Margin Clustering, MMC) と呼ばれる手法が提案されている [5]．MMC では，すべてのインスタンスがとり得るクラス分類で分離超平面を求め，その中でマージンが最大のものを，インスタンスを分離する超平面とする．

本研究では比較手法として MMC を高速に解く手法である CPMMC(Cutting Plane MMC) [5] を用いる．CPMMC では CutS3VM と同様に，ランダムに初期平面を与えその周辺のインスタンスのみを用いて新たな分離超平面を求めることで，計算量の削減を図っている．

3. 評価対象とする手法

本研究で扱う問題は，複数のキーワード情報が付与されたファイルと何も付与されていないファイルが混在するようなファイル集合からクラスタを生成することである．本研究では，次のような手法でファイルのクラスタリングを行うことを提案する（図 2）．なお本論文では，ファイル集合に含まれていると推定されるクラスの数，すなわち最終的に得られるクラスタの数を k とする．

まず，ファイル集合をキーワードがあるものとないものに分割する．次にキーワードがあるファイルについて，そのキーワードの中から各テーマを表わすものをラベル $l_1, \ldots, l_k$ として選定し，それに対応するラベル付きファイルの集合 $X_1, \ldots, X_k$ を選定する．これをラベル選定と呼ぶ．そして得られたラベルを制約として半教師付きクラスタリングを行う．本研究では，情報利得によって得られたラベルを擬似的な教師データとした CutS3VM を用い，ファイルをクラスタリングする．また CutS3VM のマルチクラスタリング化手法として 2 種類の手法を用いることを提案する（後述）．最終的には k 個のクラスタ $C_1, \ldots, C_k$ とそれに対応するラベル $l_1, \ldots, l_k$ を得る．

3.1 ラベル選定

半教師付きクラスタリングで k 個のクラスタを得るためには各クラスを表わすラベル $l_1, \ldots, l_k$ とそれに対応するラベル付きファイルの集合 $X_1, \ldots, X_k$ が必要である．そこで，このラ

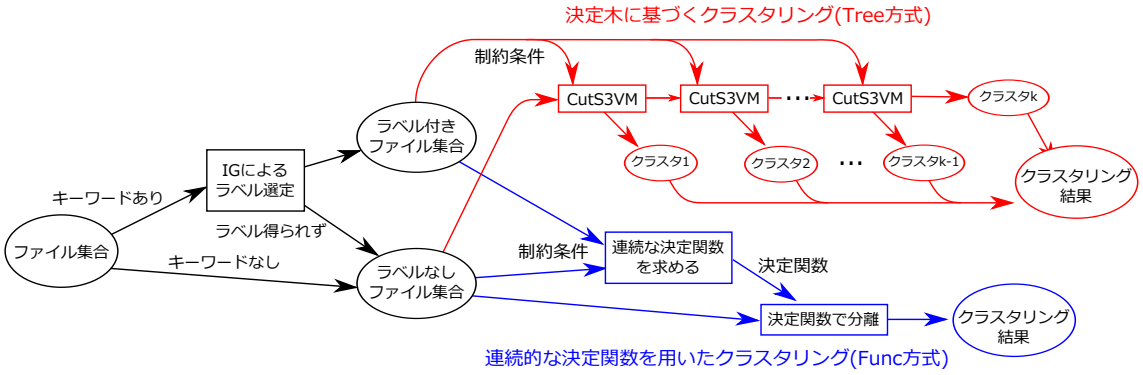

図 2: 提案手法のフロー図

ベルとファイル集合を複数のキーワード情報が付与されたファイルから推定し，ラベルが付いたファイル集合を作成する．これを本研究ではラベル選定と呼ぶ．

ラベル選定には, [2] で用いた，情報利得 (Information Gain, IG) からキーワードのスコアを求め，最も高いものを順に採用する手法を用いる．以下でその手法を説明する．

ファイル集合 X に含まれるファイルに付けられたキーワードの集合を W_X とする．最終的に得られるラベル付きファイルの各集合 $X_1, \ldots, X_k$ に含まれるファイルに付けられたキーワード集合 $W_{X1}, \ldots, W_{Xk}$ の間で共通するものが存在しなければ（$i \neq j$ のとき $W_{Xi} \cap W_{Xj} = \emptyset$ ならば），ファイル集合が適切に分離されていると考える．そこで，複数のキーワードを含むファイルの集合 χ を，なるべくこの条件を満たすようなファイル集合に分割する．

ファイル集合 χ に含まれるファイルに付与されたキーワードの 1 つ $w \in W_\chi$ に注目し，w を持つファイルの集合 X_w と w を持たないファイルの集合 $X_{\overline{w}}$ へ分割する．この分割したファイル集合 $X_w, X_{\overline{w}}$ に対して，残りのキーワード $w' \in W_\chi \setminus \{w\}$ の情報利得 $IG_w(w')$ を計算する．

$IG_w(w')$ は，キーワード w を持つインスタンスか否かという観点で分割された集合について,「キーワード w' が各集合内のインスタンスに付けられているか否か」という情報がどれだけ集合の曖昧さ（エントロピー）を減少させるかを示す値であり，キーワード w と w' が共起しやすいもしくは共起しにくい場合，$IG_w(w')$ は高い値をとる．したがって，$X_w, X_{\overline{w}}$ に対して，すべての w' について $IG_w(w')$ が高い値を示すのであれば，X_w と $X_{\overline{w}}$ に共通するキーワードが少なく，良好な状態であるといえる．

つまり，ファイル集合 χ が与えられたとき，先の条件を満たす最も良いキーワード w は，$\underset{w}{\operatorname{argmax}} \min_{w'} IG_w(w')$ により求めることができる．こうして得られたキーワード w をラベル l_1 とし，集合 X_w を l_1 に対応するラベル付きファイルの集合 X_1 とする．w を含まない集合 $X_{\overline{w}}$ に対してもこの分割手法を繰り返し適用することにより，ラベル $l_2, \ldots, l_k$ とそれに対応するラベル付きファイルの集合 $X_2, \ldots, X_k$ を選定することができる．

3.2 クラスタリング

続いて，3.1 節で得られたラベルを基にクラスタリングを行い，ラベルがついていないファイルのラベルを推定する．本研究では，情報利得によって得られたラベルを擬似的な教師データとして CutS3VM を用いファイルをクラスタリングする．しかし先述の通り CutS3VM は 2 クラスのデータを分類する手法であり，実際のデータに適用する場合にはこれをマルチクラスに拡張する必要がある．本研究では 2 種類の手法を用い CutS3VM をマルチクラスタリングに拡張することを提案する．

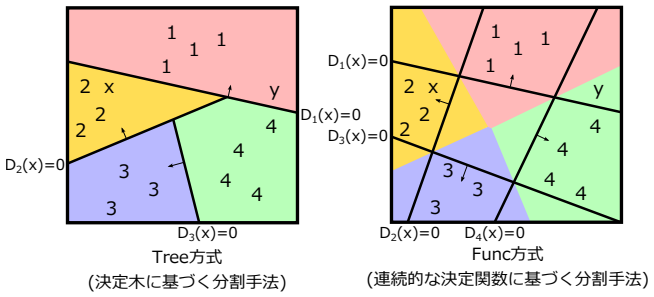

図 3: 分割手法の比較

3.2.1 Tree 方式

1 つ目は，決定木に基づき，CutS3VM を多段に適用することにより階層的に分離する手法である（図 2 上）[6]．本研究ではこれを **Tree 方式**と呼ぶ．具体的な手法は次のとおりである．

情報利得により得られたラベル付きファイルの集合を，ラベル選定で得られた順に $X_1, \ldots, X_k$ とする．ファイル集合全体に対し，X_1 を正例，$X_2 \cup \cdots \cup X_k$ を負例として CutS3VM を解く．その結果得られる決定関数を $D_1(\mathbf{x})$ として，$D_1(\mathbf{x}) > 0$ となるファイル x がラベル l_1 のクラスタに属するとする．$D_1(\mathbf{x}) < 0$ となったファイルに対しては，X_2 を正例，$X_3 \cup \cdots \cup X_k$ を負例として再帰的にクラスタリングを行うことで，ファイル集合を k 個のクラスタに分割することができる．

2 次元データに対しこの手法を適用したイメージを図 3 の左で示す．ここでは 1 から 4 のラベルが付いたインスタンスを 1 から順に 4 クラスタに分割し，インスタンス x, y の属するクラスタを判定している．直線は分離超平面 $D(\mathbf{x}) = 0$，直線から出ている矢印は決定関数の正の方向を示す．クラスタリングの結果，x はクラスタ 2 に，y はクラスタ 1 に属すると判定された．この手法の利点は，分割するごとにインスタンス数が少なくなるため高速にクラスタリングできることである．一方，ク

ラスタリング精度が分割順序に左右されることが弱点として挙げられる．

3.2.2 Func 方式

Tree 方式の問題点を解消するため，各クラスタを分離する決定関数を求めこれを連続的な決定関数に変換する手法 [7]（図 2 下）を用いることを提案する．本研究ではこれを **Func 方式**と呼ぶ．具体的な手法は次のとおりである．

情報利得により得られたラベル付きファイルの集合を，ラベル選定で得られた順に $X_1, \ldots, X_k$ とする．ファイル集合全体に対し，$i = 1, \ldots, k$ について X_i を正例，$(X_1 \cup \cdots \cup X_k) \setminus X_i$ を負例として CutS3VM を解く．得られた決定関数を $D_i(x)(i = 1, \ldots, k)$ として，ファイル x が属するクラスタを連続的な決定関数 $\underset{i=1\ldots k}{\operatorname{argmax}} D_i(x)$ で決定する．Tree 方式と比較すると，クラスタリング精度が分割順序に左右されないことがこの手法の利点である．2 次元データに対しこの手法を適用したイメージを図 3 の右で示す．x はクラスタ 2 に，y は Tree 方式と異なりクラスタ 4 に属すると判定された．

4. 評価実験

提案手法の性能を確かめるため，評価実験を行う．評価実験では，キーワード付きのファイルとして論文を用い，クラスタリングでは論文のメタ情報をベクトル化したものを用いる．

また提案手法はラベル選定とクラスタリングの 2 段階に分かれており，クラスタリング精度はラベル選定の精度に依存する．本研究ではまずラベル選定の結果を評価し，続いてクラスタリングの評価を行う．

4.1 評価実験環境

4.1.1 実験データ

評価実験では，論文データベース CiNii Articles [8] から「渡辺陽介」「横田治夫」を著者として含む論文のタイトル，著者，発表年，引用論文，キーワードを収集した．また，科学研究費補助金のデータベースである KAKEN [9] からプロジェクトおよび掲載された論文の情報を収集した．なお，同姓同名の別人による論文や明らかな誤記は手作業で除去や修正を行った．

4.1.2 論文のベクトル化

評価実験で用いるメタ情報は，論文のタイトル，著者，論文に付けられたキーワード，引用している論文，関連プロジェクト(注1)，発表年である．

前処理として，タイトルを形態素ごとに分割する．形態素解析エンジンには MeCab [10] を用いた．MeCab では，与えた文から形態素やその品詞などの情報を得ることができる．この解析結果から，名詞とアルファベット以外の形態素を削除する．またタイトル内で重複する形態素は 1 つだけ残す．接頭詞や接尾詞に分類された形態素も削除する．

論文 d_x は次のようなベクトル v_x で表現される．

$$v_x = \{t_1, \ldots, t_{n_t}, a_1, \ldots, a_{n_a}, w_1, \ldots, w_{n_w}, c_1, \ldots, c_{n_c}, p_1, \ldots, p_{n_p}, y\}$$

(注1)：その論文が研究成果報告書に掲載された，科学研究費補助金（科研費）の研究課題

表 1: 評価論文集合 1（著者名「渡辺陽介」）

No.	研究テーマ名	論文数	キーワードあり
1	配信型情報源統合	6	2
2	連続的問合せ（データストリーム）	10	5
3	分散・高信頼化（データストリーム）	10	7
4	統合環境（データストリーム）	6	2
5	テロップ認識	4	3
6	アクセス履歴からのファイル関連分析	6	3

表 2: 評価論文集合 2（著者名「横田治夫」）

No.	研究テーマ名	論文数	キーワードあり
1	負荷分散	40	27
2	自律ディスク	28	27
3	Fat-Btree	26	18
4	e-ラーニング	27	22
5	web	19	10
6	アクティブデータベース	11	7
7	冗長ディスクアレイ	7	4
8	XML	5	3
9	リサーチマイニング	5	2

ここで，t_i はタイトルの各形態素が全論文で出現する頻度とする．同様に a_i は著者の，w_i はキーワードの，c_i は引用論文の，p_i は関連プロジェクトの出現頻度とする．また y は最も古い論文の出版年 Y_0 を論文 d_x の出版年から引いた値とする．

4.1.3 評価論文集合

各論文集合を人手により分類し，これを評価論文集合とする．「渡辺陽介」を著者として含む論文を人手で分類したところ，42 本の論文を 6 個の研究テーマに分けることができた．これを評価論文集合 1 とする（表 1）．また「横田治夫」を著者として含む論文については，169 本の論文を 9 個の研究テーマに分けることができた．これを評価論文集合 2 とする（表 2）．これらの評価論文集合にどれだけ近いラベルやクラスタを求められたかで提案手法の評価を行う．

4.2 実験手法

4.2.1 評価尺度

ラベル選定の評価では，使用した論文集合に含まれるあるテーマの論文をほかのテーマの論文から分離することが可能かどうかで評価する．ラベル選定ではテーマ名とは異なるがそのテーマでしか使用されていないキーワードが得られることがある．このようなキーワードでもそのテーマの論文をほかのテーマの論文から分離することは可能なため，適切なラベルが選定されていると判定する．後述のラベル選定結果における，表 3 の「Virtual Machine Technology」や表 4 の「アクセスログ解析」などがこれに該当する．

クラスタリングの評価尺度として，エントロピー (Entropy) および純度 (Purity) を用いた．テーマ別に分類した入力論文集合を $A_1, \ldots, A_k$，クラスタを $C_1, \ldots, C_k$ とすると，それぞれの定義は以下の通りになる．

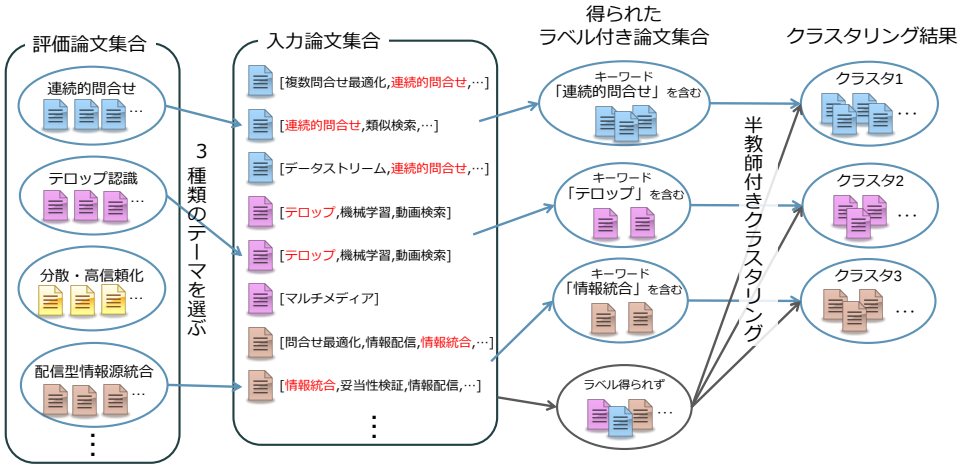

図 4: 3 クラス分類の実験手順

$$Entropy = \frac{1}{n}\sum_{r=1}^{k}|C_r|\left(-\frac{1}{\log k}\sum_{i=1}^{k}\frac{|C_r \cap A_i|}{|C_r|}\log\frac{|C_r \cap A_i|}{|C_r|}\right)$$

$$Purity = \frac{1}{n}\sum_{r=1}^{k}\max_{i}|C_r \cap A_i|$$

ここで，n は総論文数，$|A|$ は集合 A に含まれる論文数を表す．エントロピーはクラスタ内でのテーマの混ざり具合を表し，値が小さいほどクラスタリング精度が高い．また純度はクラスタに最も多く含まれる論文のテーマの割合を表し，値が大きいほど精度が高いといえる．

CutS3VM の結果はランダムな初期平面に依存するため，初期平面をランダムに変えて 10 回試行し，2 標本検定で比較手法に対する優位性を示す．手法 A と手法 B のエントロピーの母平均をそれぞれ μ_A, μ_B とし，左片側検定（有意水準 5%．以下同じ）で帰無仮説 $\mu_A = \mu_B$ を検定する．また純度についても右片側検定で同様に検定する．手法 A のエントロピーが有意に低く（左片側対立仮説 $\mu_A < \mu_B$ を採択），純度が有意に高いとき，手法 A の精度が手法 B より有意に高いとする．

4.2.2 実験手順

本手法のマルチクラス分類の性能を確かめるため，図 4 のような手法で 3 クラス分類の実験を行う．本実験では，評価論文集合から 3 種類の研究テーマを選び，それらに含まれる論文を混在させた入力論文集合を作る．この入力論文集合から，情報利得を用いてラベルとそれに対応する論文集合を 3 つ選定する．これを擬似教師データとして半教師付きクラスタリングを行い，得られたクラスタと評価論文集合からエントロピーおよび純度を求める．

これを，評価論文集合から 3 つの研究テーマを選ぶすべての組み合わせ（${}_{\text{テーマ数}}C_3$ 通り）に対して実行し，全組み合わせでの平均エントロピーと純度や，組み合わせごとのエントロピーと純度で評価を行う．

4.2.3 比較対象

比較実験として，次のような比較対象を用いた実験を行う．

a） CPMMC

情報利得で得られた擬似教師データがクラスタリングに与える効果を確認するため，教師データを全く使わない CPMMC をマルチクラスタリングに拡張した手法 [11] との比較を行う．4.2.2 節と同様にして作成した論文集合を，CPMMC で 2 クラスタに分割する．このうち論文数が多いほうにさらに CPMMC を適用し，合計 3 個のクラスタを得る．得られたクラスタと評価論文集合からエントロピーおよび純度を求め評価する．

b） 情報利得+COP-KMEANS

クラスタリングに CutS3VM を使用することによる効果を確認するため，k-means 法 [12] を半教師付きクラスタリングに応用した COP-KMEANS [13] との比較を行う．k-means 法 [12] は，ランダムに生成した初期クラスタを与え，中心が最も近いクラスタに再配分することで最適なクラスタを生成するクラスタリング手法である．これに対し，インスタンスとクラスタの関係に対する制約を加えてクラスタリングを行う COP-KMEANS [13] が提案されている．

本実験では，初期クラスタをランダムに生成する代わりに，情報利得によるラベル抽出結果から，上位 $k(=3)$ 個のラベルをキーワードとして含む論文を核となる論文として与え，それ以外の論文はクラスタ内に含まれる論文との距離の平均が最も短いクラスタに配分する．また論文を中心が最も近いクラスタに再配分する際も，核となる論文は再配分せず，クラスタとラベルの関係が保たれるようにする．得られたクラスタと評価論文集合からエントロピーおよび純度を求め評価する．なお初期クラスタをランダムに生成していないため，結果はランダム性に依存しない．

4.3 実験結果と考察

4.3.1 ラベル選定

まずラベル選定で得られたラベルを評価する．

評価論文集合 1 の 6 テーマから 3 テーマを選ぶ全ての組み合わせ (20 組) でラベル選定を行った．選んだテーマと得られたラベルとの関係は表 3 の通りである．紙幅の都合上，どのテーマにも対応しないラベルや同一のテーマに対して複数得られたラベルは省略した．空白の箇所は対応するラベルを得られなかったことを意味する．選定の結果，3 つの研究テーマに対応するラベルをすべて得られたのは表 3 に網掛けで示す 12 組

表 3: ラベル選定結果（評価論文集合 1）

テーマ 1	得られたラベル	テーマ 2	得られたラベル	テーマ 3	得られたラベル
配信型情報源統合		連続的問合せ	連続的問合せ	分散・高信頼化	VMT
配信型情報源統合		連続的問合せ	連続的問合せ	統合環境	
配信型情報源統合	情報統合	連続的問合せ	連続的問合せ	テロップ認識	テロップ
配信型情報源統合	情報統合	連続的問合せ	連続的問合せ	ファイル関連分析	データベース工学
配信型情報源統合	情報配信	分散・高信頼化	VMT	統合環境	
配信型情報源統合	情報統合	分散・高信頼化	VMT	テロップ認識	テロップ
配信型情報源統合	情報統合	分散・高信頼化	VMT	ファイル関連分析	データベース工学
配信型情報源統合	情報統合	統合環境	data stream	テロップ認識	テロップ
配信型情報源統合	情報統合	統合環境	data stream	ファイル関連分析	データベース工学
配信型情報源統合	情報統合	テロップ認識	テロップ	ファイル関連分析	データベース工学
連続的問合せ	連続的問合せ	分散・高信頼化	VMT	統合環境	
連続的問合せ	連続的問合せ	分散・高信頼化	VMT	テロップ認識	テロップ
連続的問合せ	連続的問合せ	分散・高信頼化	VMT	ファイル関連分析	データベース工学
連続的問合せ	データストリーム	統合環境		テロップ認識	テロップ
連続的問合せ	データストリーム	統合環境		ファイル関連分析	データベース工学
連続的問合せ	continuous query	テロップ認識	テロップ	ファイル関連分析	データベース工学
分散・高信頼化	データストリーム	統合環境		テロップ認識	テロップ
分散・高信頼化	データストリーム	統合環境		ファイル関連分析	データベース工学
分散・高信頼化	VMT	テロップ認識	テロップ	ファイル関連分析	データベース工学
統合環境	data stream	テロップ認識	テロップ	ファイル関連分析	データベース工学

※ VMT = Virtual Machine Technology

表 4: ラベル選定結果（評価論文集合 2，研究テーマが正しく得られたもののみ掲載）

テーマ 1	得られたラベル	テーマ 2	得られたラベル	テーマ 3	得られたラベル
負荷分散	偏り制御	自律ディスク	ネットワーク接続ディスク	web	アクセスログ解析
負荷分散	偏り制御	自律ディスク	ネットワーク接続ディスク	アクティブ DB	アクティブデータベース
負荷分散	Parallel Disk	自律ディスク	ネットワーク接続ディスク	冗長ディスクアレイ	RAID
自律ディスク	autonomous disks	e-ラーニング	教育コンテンツ	web	アクセスログ解析
web	Web	アクティブ DB	Web とインターネット	冗長ディスクアレイ	RAID
web	Web	冗長ディスクアレイ	RAID	リサーチマイニング	クラスタリング
アクティブ DB	ワークフロー	冗長ディスクアレイ	RAID	XML	XML
アクティブ DB	ワークフロー	冗長ディスクアレイ	RAID	リサーチマイニング	情報分析
アクティブ DB	ワークフロー	XML	XML	リサーチマイニング	情報分析
冗長ディスクアレイ	RAID	XML	更新	リサーチマイニング	情報分析

だった．また 1 組を除いては，2 つ以上の研究テーマに対応するラベルを得られている．このことから，評価論文集合 1 に対してはこのラベル選定手法が有効であると考えられる．

続いて評価論文集合 2 の 9 テーマから 3 テーマを選ぶ全ての組み合わせ (84 組) でラベル選定を行った．ラベル選定で得られたラベル 3 種について，表 4 に示す 10 組では 3 つの研究テーマに対応するラベルをすべて得られた．一方 84 組中 23 組では 3 つのラベルがいずれも同一の研究テーマに対応するものであった．

評価論文集合 2 で正しく得られたラベルの割合が評価論文集合 1 に比べて少ないのは，同一テーマの論文がすべて同じキーワードを持っているわけではないためと推測される．ファイル集合 χ からラベルを選定するとき，χ にキーワード付きファイル数が比較的多いテーマ A が含まれていたとする．3.1 節の手法に基づき χ をキーワード w を持つファイルの集合 X_w と持たないファイルの集合 $X_{\overline{w}}$ へ分割する際，w が A に含まれる論文の一部にしか付いていない場合は，分割を行ってもやはり A に含まれる論文が多数を占め，その結果 A に関するラベルが複数選定されやすくなる．このため，キーワード付き論文数がテーマによって大きく異なる集合ではキーワード付き論文数が少ないテーマに関するラベルが得られないことがある．今回の評価論文集合 2 では，キーワード付き論文数のテーマ間での比率は，3 つの研究テーマに合致するラベルをすべて得られたケースでは最大でも 6.75 倍だったのに対し，得られなかったケースでは最大 13.5 倍であった．このように評価論文集合 2 ではキーワード付き論文数の不均衡が多くの組み合わせで発生しており，評価論文集合 1 に比べて正しく得られたラベルの割合が低くなったと考えられる．

ラベル選定がうまく行われないと，クラスタリングに悪影響を及ぼすことが想定される．このため，内容が近いテーマ同士を併合したり，類義語に対応したりするなどラベル選定手法の改良が必要である．

4.3.2 マルチクラスタリング

前節で得た擬似教師データを用い，本手法のマルチクラスタリング性能を評価する．

評価論文集合 1 および 2 についてとり得る全ての組み合わせでクラスタリングを行った．エントロピーの平均と分散は表 5，純度の平均と分散は表 6 のようになった．また評価論文集合 2

表 5: クラスタリング結果のエントロピー

ラベル選定		あり				なし		あり	
クラスタリング手法		CutS3VM				CPMMC		COP-KMEANS	
		Tree 方式		Func 方式					
		平均	分散	平均	分散	平均	分散	平均	分散
評価論文集合 1	全組合せ	0.202	1.63e-2	0.212	1.74e-2	0.312	1.89e-2	0.152	2.54e-2
評価論文集合 2	全組合せ	0.283	0.719e-2	0.301	0.659e-2	0.255	0.814e-2	0.268	1.02e-2
	3 テーマ	0.206	0.976e-2	0.219	0.987e-2	0.292	0.587e-2	0.168	1.56e-2

表 6: クラスタリング結果の純度

ラベル選定		あり				なし		あり	
クラスタリング手法		CutS3VM				CPMMC		COP-KMEANS	
		Tree 方式		Func 方式					
		平均	分散	平均	分散	平均	分散	平均	分散
評価論文集合 1	全組合せ	0.806	1.67e-2	0.794	1.80e-2	0.701	1.75e-2	0.864	2.19e-2
評価論文集合 2	全組合せ	0.708	1.00e-2	0.695	1.02e-2	0.714	1.19e-2	0.722	1.54e-2
	3 テーマ	0.780	0.939e-2	0.769	1.09e-2	0.668	0.998e-2	0.819	2.46e-2

3 テーマ:3 テーマに対応するラベルを得られた組み合わせのみでのクラスタリング結果

については，3 つの研究テーマに合致するラベルを得られた 10 組のエントロピーと純度の平均と分散も求めた．

a） マルチクラスタリング方式の比較

まず，Tree 方式と Func 方式の比較を行いマルチクラスタリング方式を評価する．

評価論文集合 1 および 2 の全ての組み合わせに対して Tree 方式と Func 方式の比較を行うと，評価論文集合 1 では有意差がなかった．また評価論文集合 2 では Tree 方式のほうが有意に高かったが，3 テーマに合致するラベルを得られたケースのみで比較すると有意差はなかった（表 5,6）．Tree 方式では，最初に得られる分割が正しいものであればその後の分割の精度が低くても確実に 1 つは正しいクラスタが得られるため，ラベル選定の精度が低い場合では Func 方式より高い精度を得られたと推測される．

b） CPMMC との比較

次に，Tree 方式や Func 方式とラベル選定を行わない CPMMC との比較を行い，ラベルの選定がクラスタリングに与える影響を評価する．

評価論文集合 1 の全組み合わせでのクラスタリング結果で検定を行なうと，Tree 方式，Func 方式ともに CPMMC より精度が有意に高かった．また評価論文集合 2 の全組み合わせでの評価では，Tree 方式では有意な差がなく，Func 方式では精度が有意に低かった．しかし 3 テーマに合致するラベルを得られたケースのみで比較すると，両提案手法とも CPMMC より精度が有意に高かった（表 5,6）．組み合わせごとの検定を行うと，評価論文集合 1 については Tree 方式では 20 組中 12 組で，Func 方式では 20 組中 9 組で CPMMC より精度が有意に高かった．特に，キーワード付きの論文数が多い「連続的問合せ」「分散・高信頼化」や,「テロップ」「動画検索」など他で使われていないキーワードを持つ「テロップ認識」を含む組では精度が有意に高い組み合わせが多い．また評価論文集合 2 で各々の研究テーマに合致するラベルを得られた 10 組については，Tree 方式，Func 方式ともに 10 組中 7 組で精度が有意に高かった．

以上から，情報利得を用いてラベルを正しく得ることができた場合，ラベル選定を行わない手法に対して各提案手法が優位であることが示された．

c） COP-KMEANS との比較

最後に Tree 方式，Func 方式と組み合わせた CutS3VM と COP-KMEANS との比較を行い，クラスタリング手法による精度の差を評価する．

評価論文集合 1 および 2 の全組み合わせでのクラスタリング結果で検定を行なうと，どちらも Tree 方式は有意差がなく，Func 方式は COP-KMEANS に比べ精度が有意に低かった（表 5,6）．Tree 方式の CutS3VM と COP-KMEANS で組み合わせごとの検定を行うと，評価論文集合 1 については COP-KMEANS に比べ精度が有意に高かった組み合わせは 20 組中 2 組のみ，逆に精度が有意に低かった組み合わせは 20 組中 9 組であった．また評価論文集合 2 で各々の研究テーマに合致するラベルを得られた 10 組については，COP-KMEANS に比べ精度が有意に高い組み合わせが 4 組，逆に精度が有意に低い組み合わせが 5 組であった．

CutS3VM のほうが精度が高いケースがあったのは，各手法での距離計算の違いによるものと考えられる．COP-KMEANS の手法では，ファイル間の距離はシステム側で変動することがないため，距離尺度によってはファイル間の距離が同一テーマと異なるテーマで差がつかずクラスタリングがうまくいかないことがある．一方 CutS3VM では分離超平面が変化すると，ファイルと分離超平面の距離も変動するため，同一テーマのファイルの距離は小さく，異なるテーマの距離は大きくなるように自動で調整することができる．このため，COP-KMEANS に比べて適切な分離を行うことができる．

以上の実験結果から，COP-KMEANS のほうが有利である組み合わせも存在するが，テーマ内の距離と異なるテーマ間の距離があまり変わらない場合には，COP-KMEANS に比べて CutS3VM が有利であることが示された．

5. 関連研究

クラスタリングには，本研究で用いた S3VM や MMC をはじめとして様々な手法が提案されている．よく用いられる手法として k-means 法 [12] や凝集形階層的クラスタリング [14] がある．k-means 法では，ランダムに生成した初期クラスタを与え，その重心の計算とインスタンスの再配置を繰り返すことで最善なクラスタを生成する．凝集形階層的クラスタリングは，本研究のように 1 つのデータ集合を分割していくのではなく，距離の近いインスタンス同士を順番に結合することでクラスタを生成する．本研究では SVM をクラスタリングに応用した手法と，k-means 法を半教師付きクラスタリングに応用した手法を比較対象として用い評価した．

次に本研究での評価実験と同様に，論文をクラスタリングする手法に関する研究を挙げる．Nguyen らは，MMC を用いた論文のクラスタリングを提案している [11] [15]．クラスタリングに k-means 法を用いた場合と比較して精度が高いことがこの手法の特長だが，教師無しクラスタリングであるため利用者の意図とは異なるクラスタを生成する可能性があることが問題点として挙げられる．われわれの先行研究 [1] では，本研究と同様に情報利得を用いてラベルの選定を行い，COP-KMEANS を用いてクラスタリングを行った．クラスタリングでは，ファイルに付された各メタ情報に対し類似度を定義し，COP-KMEANS を実行する．このクラスタリング手法の問題点として，類似度の重みづけのパラメータを手動で設定しなければならないという点がある．またラベル選定手法を CutS3VM に適用した手法も提案している [2] が，CutS3VM は 2 値クラスタリングの手法であるため多値クラスタリングには直接適用できないことが問題点としてあげられる．本研究ではこの多値クラスタリングについて 2 種類の手法を用いることを提案し，その評価を行った．

6. まとめ・今後の課題

本論文では，情報利得を指標として擬似教師データを選定し，それを用いて半教師付き SVM(S3VM) を実行してクラスタリングをする手法において，S3VM をマルチクラスに拡張するため，決定木に基づき S3VM を多段に適用することにより階層的に分割する手法 (Tree 方式) と，連続的な決定関数を導入し非階層的に分割する手法 (Func 方式) を用いる 2 つの手法を提案した．

提案手法の性能を確かめるため，論文のメタ情報を用い評価実験を行った．ラベル選定の結果，分類に適したラベルをすべて得ることができた組み合わせの割合は論文集合によって差があった．分類に適したラベルをすべて得ることができた組み合わせのみを用いてクラスタリングを行うと，エントロピーと純度の 2 つの指標においてクラスタリング精度がラベル選定を行わない場合よりも有意に高かった．またマルチクラスタリング手法の比較では，Func 方式に比べ Tree 方式のほうが精度が有意に高いケースと，有意差がないケースが存在した．

今後の課題をいくつか挙げる．まず，ファイル集合によっては適切なラベルを選べない事象が確認されているため，ラベル選定精度を向上させなければならない．また，本研究ではマルチクラスタリング手法として Tree 方式と Func 方式の 2 つを比較したが，この他にもペアワイズ SVM や同時定式化 SVM と呼ばれる手法も存在する．これらの手法を用いることで精度が向上する可能性もあるため，比較実験を行う必要がある．さらに，現在は評価実験に論文のみを使用しているが，手法自体は他の種類のデータにも適用可能である．このため他の種類のデータセットに適用した実験が必要になると考えられる．

謝　辞

本研究の一部は、日本学術振興会科学研究費補助金基盤研究（A）(#22240005, #25240014) の助成により行われた。

文　献

[1] 加藤大智, Manh Cuong Nguyen, 橋本泰一, 横田治夫,「論文のラベル付きクラスタリングのための情報利得を用いたキーワード選定」, DEIM Forum 2012, E10-1, 2012.

[2] 加藤大智, 橋本泰一, 渡辺陽介, 横田治夫,「ラベル付きクラスタリングにおける情報利得を用いたキーワード選定結果の適用手法」, DEIM Forum 2013, C10-4, 2013.

[3] Vladimir N. Vapnik and A. Sterin, “On structural risk minimization or overall risk in a problem of pattern recognition”, In *Automation and Remote Control*, Vol. 10, pp. 1495–1503, 1977.

[4] Bin Zhao, Fei Wang, and Changshui Zhang, “CutS3VM: a fast semi-supervised SVM algorithm”, In *KDD ’08: Proc. of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 830–838, 2008.

[5] Bin Zhao, Fei Wang, and Changshui Zhang, “Efficient maximum margin clustering via cutting plane algorithm”, In *The 8th SIAM International Conference on Data Mining (SDM 08)*, pp. 751–762, 2008.

[6] F. Takahashi and S. Abe, “Decision-tree-based multiclass support vector machines”, In *Proc. of ICONIP ’02* , Vol. 3, pp. 1418–1422 vol.3, 2002.

[7] 阿部重夫,「パターン認識のためのサポートベクトルマシン入門」. 森北出版, 2011.

[8] 国立情報学研究所,「CiNii articles - 日本の論文をさがす」, `http://ci.nii.ac.jp/`.

[9] 国立情報学研究所,「KAKEN - 科学研究費補助金データベース」, `http://kaken.nii.ac.jp/`.

[10] “MeCab: Yet another part-of-speech and morphological analyzer”, `http://mecab.sourceforge.net/`.

[11] Manh Cuong Nguyen, Daichi Kato, Taiichi Hashimoto, and Haruo Yokota, “Research history generation using maximum margin clustering of research papers based on metainformation”, In *Proc. of iiWAS ’11*, pp. 20–27, 2011.

[12] James MacQueen, “Some methods for classification and analysis of multivariate observations”, In *Proc. of the 5th Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics*, pp. 281–297, 1967.

[13] Kiri Wagstaff, Claire Cardie, Seth Rogers, and Stefan Schrödl, “Constrained k-means clustering with background knowledge”, In *Proc. of ICML ’01*, pp. 577–584, 2001.

[14] 齋藤堯幸, 宿久洋,「関連性データの解析法 — 多次元尺度構成法とクラスター分析法」. 共立出版, 2006.

[15] Manh Cuong Nguyen, 加藤大智, 橋本泰一, 横田治夫,「論文のメタ情報を利用した研究履歴自動抽出・可視化システム」, DEIM Forum 2012, F6-3, 2012.