Naive Bayesモデルを用いた効率的なクラスタラベリング手法

Efficient cluster labeling using naive Bayes models

小島諒介[1]　亀谷由隆[2]　佐藤泰介[1]
Ryosuke KOJIMA [2]　Yoshitaka KAMEYA [2]　Taisuke SATO[1]

[1] 東京工業大学 大学院情報理工学研究科 計算工学専攻
[1] Department of Computer Science, Graduate School of Information Science and Engineering
Tokyo Institute of Technology
[2] 名城大学 理工学部 情報工学科
[2] Department of Information Engineering, Faculty of Science and Technology
Meijo University

Abstract: Cluster labeling is used to generate descriptive labels of data clusters, which are easy to read and summarize sufficiently well the clusters. In previous work, cluster labels are limited to be a conjunction of attribute-value pairs, but in practice such labels tend to be too fragmentary and resemble each other. Introducing a disjunctive normal form (DNF) of attribute-value pairs mitigates this limitation, but may cause another problem that the search space becomes larger. In this paper, we propose an efficient method for finding cluster labels in DNF and empirically demonstrate the effectiveness of our apporach.

1　はじめに

データマイニングの手法の一つにクラスタリングがある．クラスタリングとはデータ集合を内的結合と外的分離が達成されるような部分集合に分離することである．クラスタリングによって得られたクラスタ中の事例には共通した傾向が見られる．クラスタの傾向は予測等に用いることができるため価値のある情報として利用される．現実のデータにおいて，クラスタ中の事例は大量になることが多いためクラスタの傾向を掴むことは容易ではない．そこで，クラスタ中の事例の要約を自動的に行う手法が必要となる．クラスタ中の事例を要約したものはラベルと呼ばれ，ラベルを生成しユーザに提示することはクラスタラベリングと呼ばれる [1].

クラスタラベリングにおいては，データを直接見ることなく要約されたラベルのみでそのクラスタを理解できることが理想であり,「クラスタを正確に表しているか」と「出力されるラベルがユーザに理解し易いか」という 2 点から評価される．これらの評価はラベルの表現方法により大きく影響を受ける．例えば，属性値の集合や属性値の AND 式 (連言) によりラベルを表現する方法が提案されている [2]．しかし，これらの方法では経験的に似たラベルが多く見つかることが問題として挙げられる．また，クラスタによっては AND 式のみによる単純な表現ではクラスタを正確に表すことができない場合がある．

そこで本研究では，属性値の AND/OR 式で表現されるラベルを用いてクラスタラベリングを行う手法を提案する．この時，属性値の組み合わせの探索は膨大な計算になるため，AND/OR 式を DNF 式に制限し適切な枝刈りをしながら探索を行う．ラベルとして AND 式より複雑な構造の DNF 式を用いることにより，AND 式では表現できないクラスタを表現可能になる．一方，作成されたラベルはユーザにとって理解しやすいものであることが望まれる．このことから提案法では，適切かつ簡潔なラベルを作ることを目標とする．

2　AND ラベルと既存手法

ここではまず，既存手法について説明する．既存手法は属性と属性値の AND 式をラベルとし，このラベルを AND ラベルと呼ぶ. 例えば「色」「形」「大きさ」という属性があり，属性値として「色」に対して「赤」と「青」,「形」に対して「丸」と「三角」,「大きさ」に対して「大」と「小」がある時,「色=赤 ∧ 形=丸」や「色=青 ∧ 形=三角 ∧ 大きさ=大」が AND ラベルである．以降簡単のため，特に誤解がない場合は属性名を省略する. この略記により，先の例のラベルは「赤 ∧ 丸」,「青 ∧ 三角 ∧ 大」のように表される．

2.1 AND ラベルの評価基準

クラスタにふさわしい AND ラベルを評価する方法として，情報検索などでよく用いられる F 値がある．AND ラベル L_{AND} で表現されるデータがクラスタ c 中に出現する確率を再現率 $p(L_{\mathrm{AND}} \mid c)$ と呼ぶ．一方 L_{AND} を満たすデータがクラスタ c に属する確率を精度 $p(c|L_{\mathrm{AND}})$ と呼ぶ．このとき，F 値は再現率と精度の調和平均として以下のように定義される．

$$F(L_{\mathrm{AND}}) = \frac{2}{1/p(L_{\mathrm{AND}} \mid c) + 1/p(c \mid L_{\mathrm{AND}})}$$

F 値が大きいほど $(0 < F(L_{\mathrm{AND}}) \leq 1)$ その AND ラベルの表現するデータ集合とクラスタ内のデータ集合がよく一致しているとみなすことができる．

2.2 Naive Bayes モデルと AND ラベル

クラスタリングには様々な手法があるが，ここでは Naive Bayes モデルに EM アルゴリズムを適用したクラスタリング手法について取り上げる．クラスタリングに Naive Bayes モデルを用いると，Naive Bayes モデルの条件付き独立の性質によりクラスタ c における AND ラベル L_{AND} の再現率 $p(L_{\mathrm{AND}} \mid c)$ を効率的に求めることができる．例えば，$p(A_1 = a_1 \wedge A_2 = a_2 \mid c)$ という確率は条件付き独立より

$$p(A_1 = a_1 \wedge A_2 = a_2 \mid c) = p(A_1 = a_1 \mid c)p(A_2 = a_2 \mid c)$$

として計算できる．右辺の確率はそれぞれモデルのパラメータとして得られる．

また，AND ラベルの評価値である F 値を計算するためには精度 $p(c \mid L_{\mathrm{AND}})$ を計算する必要がある．$p(c \mid L_{\mathrm{AND}})$ についてはベイズの定理より

$$p(c \mid L_{\mathrm{AND}}) = \frac{p(L_{\mathrm{AND}} \mid c)p(c)}{p(L_{\mathrm{AND}})} = \frac{p(L_{\mathrm{AND}} \mid c)p(c)}{\sum_{c'} p(L_{\mathrm{AND}} \mid c')p(c')}$$

として計算できる．ただし，$p(c)$ はクラスタリング時に推定されるのでここでは既知として使用できる．このようにして任意の AND ラベルについて F 値を計算できる．

2.3 RP-growth

AND ラベルの探索方法として，F 値の大きい上位 k 個の AND ラベルを出力することを目的とする RP-growth[3] を説明する．RP-growth では以下の条件を満たす AND ラベル集合 R を生成する $(|R| = k)$．

- $\forall L_{\mathrm{AND}} \in R$ に対し $p(L_{\mathrm{AND}} \mid c) \geq s$
- $\forall L_{\mathrm{AND}} \in R$ に対し $p(c \mid L_{\mathrm{AND}}) \geq r$
- R 中のどの 2 つのラベルも「より弱い」関係 (後述) にない
- R に属さないラベルで上記条件を満たし，R 中のラベルより F 値が高いラベルは存在しない．

ただし，候補集合の大きさ $k(\geq 1)$，再現率の下限 $s(\geq 0)$，精度の下限 $r(\geq 0)$ はユーザが指定するパラメータである．

RP-growth では，冗長なラベルを出力しないという観点から「より弱い」という概念を導入している．$L'_{\mathrm{AND}} \supset L_{\mathrm{AND}}$ かつ $F(L'_{\mathrm{AND}}) \leq F(L_{\mathrm{AND}})$ の時「L'_{AND} は L_{AND} より弱い」という．

RP-growth は接尾探索木を深さ優先で効率的に探索を行う．探索中は候補集合として常に k 個の AND ラベルのみを持つようにし，新たに訪問した AND ラベルの F 値が候補集合 k 個の中で最も F 値の小さい AND ラベルよりも大きい場合には，新しい AND ラベルを加え，最も F 値の小さい AND ラベルを取り除く．このようにして探索すると，最終的に候補集合中に F 値が上位 k 個のラベルが得られる．

また，候補集合で他の AND ラベル「より弱い」AND ラベルが含まれないようにすることを考える． RP-growth の探索順序では，AND ラベル L_{AND} を新たに訪問する時，AND ラベル L_{AND} の真部分集合はすでに訪問済みである．つまり「より弱い」の定義より，新たな AND ラベル L_{AND} が他の AND ラベル L'_{AND}「より弱い」ならば L'_{AND} については既に探索が終わっているはずである．したがって，L_{AND} が他の AND ラベルより弱いかどうかを確認するには L_{AND} を訪問した時の候補集合中の AND ラベルより弱いかどうかを確認すれば十分である．

RP-growth では反単調性が成り立つ F 値の上限を求め，分枝限定法を用いることによって枝刈りを行なっている． AND ラベルにおける F 値の上限 $\bar{F}_{\mathrm{AND}}$ は $p(c \mid L_{\mathrm{AND}}) = 1$ として以下のようにして計算される。

$$\begin{aligned} F(L_{\mathrm{AND}}) &\leq \frac{2}{1/p(L_{\mathrm{AND}} \mid c) + 1} \\ &= \bar{F}_{\mathrm{AND}}(L_{\mathrm{AND}}) \end{aligned}$$

この上限 $\bar{F}_{\mathrm{AND}}(L_{\mathrm{AND}})$ については反単調性が成り立ち，RP-growth の探索順から枝刈りを行うことができる．また，同様に F 値の上限を用いて「より弱い」関係についての枝刈りも行うことができる．

2.4 従来法の問題点

RP-growth は,「より弱い」関係を用いて冗長なラベルを出力しない．しかし，この方法を用いたとしても似

たラベルが多く出力されることがある．それは,「赤 ∧ 円」,「赤 ∧ 大」,「赤 ∧ 小」のようなラベルである．これらのラベルは属性値「赤」が全てに共通して出てくるが，これらはお互いに部分集合の関係にはない AND ラベルであるため,「より弱い」関係で簡略化することはできない．こういったラベルが大量に出力されると，重要なラベルが埋もれてしまう恐れがある．また，共通して出てくる属性値 (この場合は「赤」) は他の属性値と比較して重要な属性値であると考えられる．しかし，従来法では共通の属性を他の属性値と同等に扱ってしまっている．したがって，共通の属性を考慮したラベルの表現とラベルの探索が求められる．

3 提案法:AND/OR 式によるクラスタラベリング

我々は前節の議論に基づきラベルを属性と属性値の組から成る AND/OR 式で表現する手法を提案する．この AND/OR 式で表現されたラベルを以降 AND/OR ラベルと呼ぶ. 例えば,「(色=青 ∧ 形=三角) ∨ (大きさ=大)」のように AND/OR ラベルを表現する．この時，属性と属性値の組の数を AND/OR ラベルの長さと呼ぶことにする．以降簡単のため，特に誤解がない場合は属性名を省略する．

提案法では次のような AND/OR ラベルの探索・提示を目的とする．

- ラベルの式が複雑でない
- ラベルが冗長でない
- ラベルの表現するデータとクラスタ内のデータがよく一致している
- ユーザに複数のラベルを提示する
- 出力した複数ラベルがクラスタ内のデータをほぼ被覆している．

3.1 DNF ラベル

クラスタを表すラベルとして，任意の AND/OR 式を許すと式が複雑になる．また，同じデータを表すラベルに複数の表現ができるため探索が難しくなる．そこで探索時には，ラベルに用いる AND/OR 式を DNF 式 (属性値の連言節の選言) に制限する．以降，この DNF 式によるラベルを DNF ラベルと呼ぶ.

DNF ラベルは，AND ラベルの式を選言で結合したものとみなせる．また，選言は順序を考慮しないことから DNF ラベルは AND ラベルからなる集合とみなせる．以降では DNF ラベル L_{DNF} と書いて論理式 L_{DNF} を表すと同時に，AND ラベルからなる集合のように扱うが適宜読み替えることとする．

3.2 手法概要

提案法では以下の手順でクラスタを特徴付ける AND/OR ラベルを作成する．

1. Naive Bayes モデルを用いてクラスタリングを行いクラスタとパラメータを得る (2.2 節).
2. 各クラスタについてラベルを作成する
   (a) 属性値の組み合わせによって複数の AND ラベルを作成する
   (b) 得られた AND ラベルの組み合わせによって複数の DNF ラベルを作成する
3. 後処理をしてラベルをユーザに提示する

提案法では有効な DNF ラベルは有効な AND ラベルから作られるという仮定の下に，まず評価値の高い AND ラベル複数作成し，それらの組み合わせにより DNF ラベルを作成する．

Naive Bayes モデルを用いたクラスタリングと AND ラベルの作成については既に説明したので，以降の節では DNF ラベルの探索と後処理・表示方法について説明する．また，複数のラベルを提示する場合の手法についても説明する．

3.3 DNF ラベル探索

AND ラベルを AND/OR ラベルに一般化したとしても評価値として F 値を用いる議論は同様に成り立つ．2.3 節では属性値の組み合わせによって AND ラベルを探索する方法を示した．ここでは，DNF ラベルを AND ラベルの組み合わせによって 2.3 節と同様の方法で探索できることを示す.

RP-growth における「より弱い」という関係は $L'_{\mathrm{DNF}} \supset L_{\mathrm{DNF}}$ かつ $F(L'_{\mathrm{DNF}}) \leq F(L_{\mathrm{DNF}})$ の時「L'_{DNF} は L_{DNF} より弱い」とすることで拡張できる．DNF ラベルにおける RP-growth でも AND ラベル同様の探索順序を用いることで効率的に探索できる．DNF ラベルにおいて AND ラベルの場合と唯一異なるのは枝刈りの方法である．AND ラベルでは $p(c \mid L_{\mathrm{AND}}) = 1$ とすることで反単調性が成り立つ F 値の上限 $\bar{F}_{\mathrm{AND}}(L_{\mathrm{AND}})$ を得ていた．しかし，DNF ラベルでは同様の式では単調性を満たさない．そこで，$p(L_{\mathrm{DNF}} \mid c) = 1$ と仮定して $p(c \mid L_{\mathrm{DNF}})$ の上限を次のように求める．

$$
\begin{aligned}
p(c \mid L_{\mathrm{DNF}}) &= \frac{p(L_{\mathrm{DNF}}|c)p(c)}{p(L_{\mathrm{DNF}}|c)p(c) + p(\neg c,\, L_{\mathrm{DNF}})} \\
&\leq \frac{p(c)}{p(c) + p(\neg c,\, L_{\mathrm{DNF}})}
\end{aligned}
$$

これを用いると，新たに反単調性が成り立つ F 値の上限 $\bar{F}_{\mathrm{DNF}}(L_{\mathrm{DNF}})$ が次のようにして求まる．

$$
\begin{aligned}
F(L_{\mathrm{DNF}}) &\leq \frac{2}{1 + 1/p(c \mid L_{\mathrm{DNF}})} \\
&\leq \frac{2}{1 + \frac{p(c)+p(\neg c,\, L_{\mathrm{DNF}})}{p(c)}} \\
&= \frac{2p(c)}{2p(c) + p(\neg c,\, L_{\mathrm{DNF}})} \\
&= \bar{F}_{\mathrm{DNF}}(L_{\mathrm{DNF}})
\end{aligned}
$$

この新たな上限 $\bar{F}_{\mathrm{DNF}}(L_{\mathrm{DNF}})$ は反単調性を満たすので，RP-growth 同様の枝刈りを行うことができる．

3.4 後処理・ラベルの表示

AND ラベルでは経験的に「赤 ∧ 円」,「赤 ∧ 大」のように共通した属性値を持つラベルが多く見つかる．DNF ラベルでは，このような AND ラベルによって構成される場合「赤 ∧ (円 ∨ 大)」のようにして共通項をまとめることができる．ここでは，DNF ラベルを構成するすべての AND ラベルが共通項を持つ場合に限りそれらをまとめることとする．より複雑な論理式を許すことで，より短い AND/OR ラベルを作成することが可能だが，ユーザがその AND/OR ラベルを理解することが難しくなるためここでは共通項をまとめるという操作のみに留める．

探索時に積極的に共通項を持つ DNF ラベルを得るためには共通項の数の下限に関するパラメータ i_{DNF} を導入する ($i_{\mathrm{DNF}} \geq 0$)．i_{DNF} より共通項の数が少ない DNF ラベルを枝刈りすることで、i_{DNF} 以上の共通項をもつ DNF ラベルのみを効率的に得られる．

次に，このようにして共通項をくくりだした AND/OR ラベルの表示方法について説明する．AND/OR ラベルは 2 列からなる表にとして表示する．A，B，C，... をそれぞれ AND ラベルとして 「A ∧ (B ∨ C∨ ...)」という形式でラベルが作成されたとする．AND ラベルは「属性=属性値」を要素とした集合の形式 (中括弧) で表現し，表の 1 列目に A を，2 列目に B，C，... を列挙する．例えば「色=赤 ∧ (形=円 ∧ 大きさ=小 ∨ 形=四角)」というラベルについては表 1 のようにして表示する．なお，ラベル中の要素の順番は要素単体での F 値が高い順に並べることとする．

表 1: ラベル表示の例
「色=赤 ∧ (形=円 ∧ 大きさ=小 ∨ 形=四角)」

L	
{ 色 = 赤 }	{ 形 = 円,大きさ = 小 } { 形 = 四角 }

3.5 複数の DNF ラベルの提示

クラスタリングした後に，データを直接見ることなく要約されたラベルのみでそのクラスタを理解できることが理想である．そのため，規則学習における被覆 (covering) 法に類似した方法を用いて，指定された数の DNF ラベルでクラスタ内のデータを多く被覆する方法を採用する．

この複数 DNF ラベルを取得するアルゴリズムの擬似コードを Algorithm 1 に示す．このアルゴリズムではクラスタごとに t 個の DNF ラベルを出力する．1 行目で RP-growth により AND ラベル集合を作成している．4 行目ではその組み合わせから DNF ラベルを作っている．そして，5，6 行目で一度 DNF ラベルで利用した AND ラベルを元の AND ラベル集合から取り除いている．この 3 行目以降の DNF ラベルの操作を t 回繰り返すことで，クラスタ内のデータを多く被覆するように複数ラベルを得ることができる．

Algorithm 1 LabelGen(t)

Require: t:出力するラベル数

1: S := RP-growth によって得られた AND ラベル集合
2: $R := \emptyset$
3: **for** 1 **to** t **do**
4: L := 拡張した RP-growth によって S から DNF ラベルを一つ得る ($k = 1$)
5: L を R に追加
6: $S := S \setminus L$
7: **end for**
8: **return** R

4 評価実験

評価データとして UCI ML Repository から zoo データセット [4]，また UCI KDD Archive から 20 newsgroup データセット [5] を NaiveBayes モデルを用いてクラスタリングし，AND ラベルのみの方法と提案法の比較を行う．Naive Bayes モデルの学習には EM アルゴリズムを用い，再出発回数は 100 回，クラスタ数は真

のクラス数とした．各データセットで明瞭なクラスタを一つずつ選び，従来法 (RP-growth) で得られた AND ラベル上位 10 個，および提案法で得られた DNF ラベル上位 3 個を示す．なお，実験は C++言語による実装で，CPU が Core i7 Quad 3.40GHz のマシンで行った．

zoo データセットは事例数が 101，属性数が 16 の 7 クラスデータである．zoo データセットは動物が事例になっており，動物の特徴が属性である．属性「足」に関しては足の数を値としてとり，残りは「あり (T)」，「なし (F)」の 2 値を取る．

提案手法はパラメータによって得られるラベルが異なるが，ここでは現実的な時間で探索が終わり，かつ解釈が容易になるようパラメータを調整した．AND ラベル探索における候補集合の大きさ k は 10 とし，再現率の下限 s，精度の下限 r はともに 0.1 とした．DNF ラベル探索におけるパラメータは $s = 0$，$r = 0$，共通項数の下限 i_{DNF} は 1 とした．この時，AND ラベルの探索に 0.88 秒，DNF ラベルの探索に 0.12 秒かかった．この結果を表 2 に示す．従来法でも提案法でも「授乳かつ水棲」が上位にあることが分かる．しかし，提案法では「授乳」を共通項としてくくり出すことができ，「水棲」ということ以外に,「ひれがある」,「足がない」といった水棲の哺乳類の特徴が明確に発見できる．これは実際に事例数 6 の水棲の哺乳類のクラスタであった．

表 2: zoo データに対し従来法で得られた AND ラベル（左）と提案法で得られた AND/OR ラベル（右）

L	$F(L)$
{ 授乳=T, 水棲=T}	0.999
{ 卵=F, 水棲=T}	0.832
{ 授乳=T, ひれ=T}	0.800
{ 卵=F, ひれ=T}	0.714
{ 授乳=T, 足=0}	0.666
{ 水棲=T, 猫サイズ=T}	0.626
{ ひれ=T, 猫サイズ=T}	0.571
{ 卵=F, 足=0}	0.549
{ 足=0, 猫サイズ=T}	0.441
{ 授乳=T, 捕食=T}	0.428

L		$F(L)$
{ 授乳=T}	{ 水棲=T} { ひれ=T} { 足=0}	0.999
{ 卵=F}	{ 水棲=T} { ひれ=T}	0.832
{ 猫サイズ=T}	{ ひれ=T} { 足=0}	0.641

属性数・クラス数の多い大規模の実データとして，20 news group データ [5] を用いる．このデータは 20 トピック 18846 事例のドキュメントからなる．今回の実験ではドキュメント中に出現する単語を属性とし，テキスト中に単語が存在するかしないかの 2 値として取り扱う．ただし，本実験では文献 [2] と同様の前処理をしたものを使用する．この前処理を行った結果，属性数 (単語数) は 5129，事例数は 17930 となった．また，単語が出現しないことを表すラベルは探索から除外した．これは経験的に「文書に “〜” という単語が出現しない」という情報は有用な知見の発見にはつながりにくいからである．

この実験は，AND ラベル探索のパラメータは $k = 10$，$s = 0$，$r = 0$ とし，DNF ラベル探索のパラメータは $s = 0$，$r = 0$，$i_{DNF} = 0$ とした．この時，AND ラベルの探索に 19.59 秒，DNF ラベルの探索に 177.43 秒かかった．この結果を表 3 に示す．従来法では何度も「god」という属性が出てきているのがわかる．提案法ではそれらを共通項としてくくり出すことができている．また，提案法により F 値の高いラベルが生成されていることもわかる．実際に，このクラスタは 512 事例からなるクラスタで，そのうち宗教関連のトピック (「soc.religion.christian」と「talk.religion.misc」) が 333 事例あった．

表 3: 20 news group データに対し従来法で得られた AND ラベル（左）と提案法で得られた AND/OR ラベル（右）

L	$F(L)$
{peopl,god}	0.505
{god,time}	0.478
{god,thing}	0.465
{god,make}	0.458
{god,point}	0.445
{god,don}	0.430
{god,fact}	0.421
{god,person}	0.421
{god,part}	0.411
{god,read}	0.411

L		$F(L)$
{god}	{peopl} {part}	0.526
{god}	{fact} {person} {read}	0.504
{god}	{time} {thing}	0.499

5 関連研究

既存研究としてクラスタリングした結果に対してラベル付けを行うという手法には，トピックモデルに基づくクラスタラベリング [6] や自己組織化マップ (SOM) に対してラベル付けを行う LabelSOM [7] などがある．他の方法については Treeratpituk と Callan の論文 [8] にも記述がある．これらの方法ではラベルの長さを制限し，ラベルは AND 式のみを扱っている．

クラスタリングとラベル付けを同時に行う手法として CLIQUE [9] がある．CLIQUE は結果が DNF 表現になるという点で提案手法と共通している．他に概念クラスタリングの手法として知られる COBWEB [10] はクラスタリングとそのクラスタを記述する表現を同時に求められる．COBWEB はオンライン学習であると共にデータの入力順に依存するという性質をもつ．これらの方法と比べ提案法では NaiveBayes モデルを利用するため，クラスタ数選択のためのモデル選択スコアや欠測値を自然に扱うことができる．また，データの入力順に依存しない．

クラスタングには C4.5 などの決定木を用いる方法がある．決定木を用いる方法ではあるクラスタを表すのに冗長な表現がされる場合がある．例えば,「色=赤」で

ある事例からなるクラスタ c_1 と,「形=三角」である事例からなるクラスタ c_2 がある場合を考える. 決定木の根が「色」であるとすると、「色=赤」と「色 $\neq$ 赤」に分離され，c_1 は「色=赤」の結果としてとり出される. しかし、「形」が「色 $\neq$ 赤」の次のノードであったとしても，c_2 は「色 $\neq$ 赤 $\wedge$ 形=三角」の結果としてとり出される. この場合では c_2 を特徴付けるには「形=三角」で十分であるにもかかわらず冗長な記述となる.

評価値の上界を用いて同様に分枝限定法を用いる方法として AprioriSMP [11] があるが，この方法では「より弱い」という関係を用いていないので冗長なラベルを生成しがちである. また，提案法では評価値の上界を AND ラベルの探索としてだけではなく，DNF ラベルの探索時にも上界を用いて分枝限定法を行なっている.

6 結論・考察

本研究では，Naive Bayes モデルを用いたクラスタリングにおいて，AND/OR 式を用いたクラスタラベリングの手法を提案した. 提案法を用いて実験を行い，従来法の問題点であった共通項を持ったラベルが大量に見つかることについても，共通項をまとめることによってより簡潔なラベルが得られることがわかった.

既にクラスタリングされたデータに AND ラベルを付ける場合には，各クラスタ内の事例にパターン発見手法を直接適用することが考えられる. また，既にクラスタリングされたデータに DNF ラベルを付ける場合に提案手法を適用するには Naive Bayes モデルのパラメータを完全データから計算すればよい. また，連続値をもつデータへの対応は文献 [2] と同様に離散化を行うことで対応できる.

提案法では F 値を Naive Bayes モデルパラメータを用いることで AND ラベルと同様に計算しているが，データから直接計算する場合，効率の良いデータのスキャン方法を別途考える必要がある. また，提案法では F 値を用いているがその他の評価値でも同様にアルゴリズムを構成できる場合がある [3].

問題点として，パラメータによってはラベル探索に膨大な時間がかかる点が挙げられる. このパラメータの決定方法は現在、試行錯誤による部分が大きい. そのため、パラメータの決定またはパラメータに大きく依存しないアルゴリズムの開発は今後の課題である. また，Naive Bayes モデルというモデルが単純なため，クラスタリングの段階でうまくクラスタに分けられないことがある. より複雑なモデルに対しても同様のアプローチにより有効なクラスタラベリングの手法が必要であると考えられる.

参考文献

[1] Manning, C. D., Raghavan, P. and Schutze, H.: *Introduction to Information Retrieval*, Cambridge University Press (2008).

[2] Kameya, Y., Nakamura, S., Iwasaki, T. and Sato, T.: Verbal characterization of probabilistic clusters using minimal discriminative propositions, *Proc. of the 23rd IEEE Int'l Conf. on Tools with Artificial Intelligence*, pp. 873–875 (2011).

[3] Kameya, Y. and Sato, T.: RP-growth: Top-k mining of relevant patterns with minimum support raising, *Proc. of the 2012 SIAM Int'l Conf. on Data Mining* (2012).

[4] Frank, A. and Asuncion, A.: UCI Machine Learning Repository (2010).

[5] Lang, K.: NewsWeeder: Learning to filter netnews, in *Proc. of the 12th Int'l Machine Learning Conf.*, pp. 331–339 (1995).

[6] Mei, Q., Shen, X. and Zhai, C.: Automatic labeling of multinomial topic models, in *Proc. of the 13th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining*, pp. 490–499 (2007).

[7] Rauber, A.: LabelSOM: On the labeling of self-organizing maps, in *Proc. of Int'l Joint Conf. on Neural Networks*, pp. 1–6 (1999).

[8] Treeratpituk, P. and Callan, J.: Automatically labeling hierarchical clusters, in *Proc. of the 2006 Int'l Conf. on Digital Government Research*, pp. 167–176 (2006).

[9] Agrawal, R., Gehrke, J., Gunopulos, D. and Raghavan, P.: Automatic subspace clustering of high dimensional data for data mining applications, in *Proc. of the 1998 ACM SIGMOD Int'l Conf. on Management of Data*, SIGMOD '98, pp. 94–105 (1998).

[10] Fisher, D. H.: Knowledge acquisition via incremental conceptual clustering, *Machine Learning* (1987).

[11] Morishita, S. and Sese, J.: Transversing itemset lattices with statistical metric pruning, in *Proc. of the 19th ACM SIGMOD-SIGACT-SIGART symposium on Principles of Database Systems*, pp. 226–236 (2000).