論　文

教師付き学習を用いた教師無し変化解析手法

松澤　裕史[†]　　比戸　将平[††]　　井手　剛[††]　　鹿島　久嗣[†††]

Unsupervised Change Analysis using Supervised Learning

Hirofumi MATSUZAWA[†], Shohei HIDO[††], Tsuyoshi IDÉ[††], and Hisashi KASHIMA[†††]

あらまし　本論文では，異なる 2 つのデータに内在する差異を分析するという実用的に重要な問題（変化解析問題）を新たに定義し，その解法を示す．変化解析問題のゴールは，2 つのデータの間の何らかの差異の存在を前提に，両者の違いに対して説明を与えることである．差異の存在自体というよりは，差異についての，内在的な自由度による詳細な説明を求めるという点で，変化検出とは異なる．変化解析は，直接的な訓練データが与えられないという点において，本質的に教師なし学習の問題である．しかし本論文では，教師付き学習器を使ってこの問題が解けることを示す．核となるアイディアは，データの相違を，教師付き分類器を使って評価する点である．論文では，実データを用いて実験を行い，提案手法の有用性を示す．

キーワード　変化解析, 2 標本検定, コンセプトドリフト．

1. はじめに

外れ値検出問題は，教師無し学習の典型的な問題のひとつである．これは，あるサンプルに着目して，それと残りのデータとの間の何らかの距離尺度に基づいて，そのサンプルがどの程度異常かを決定する問題である．外れ値検出と似た問題に，変化検出という問題もある．典型的にはこれは，ある時系列的なデータを前提にして，データの確率分布の変化についての統計的検定の問題として定式化される．

しかしながら，実用上の多くの問題において，変化検出は，単なる統計的検定とは多少異なる目的意識で行われることが多い．例えば，マーケティングの用途のために，過去の顧客リストと直近の顧客リストを比べて，客層の変化を調べることがしばしば必要となる．ここでは，変化が存在すること自体はあまり問題にはならず（なぜなら過去と直近で変化がまったくないことは考えにくい），「どのように」変化が生じているのかを把握することの方が重要である．

本論文では，この実用的で重要な問題を定式化する．我々は，これを**変化解析**問題と呼ぶ．変化検出とは異なり，我々の関心は，2 つのデータ集合間にある変化を一般的に記述する枠組みを実現することにある．明らかに，変化解析問題は本来教師無し学習の問題である．我々の基本的な想定として，2 つのデータ集合が，ベクトルの集合として与えられていると考える．我々の目的は，データ集合の発生源の詳細やデータ構造に関する内部情報を与えられることなしに，2 つのデータ集合の比較に基づいて，差異の起因になる情報を発見することである．本論文の主たる貢献は，本質的に教師無し学習の問題である変化解析問題が，**教師付き学習**によって解決されることを示した点にある．

2 つのデータ集合を比較する問題は，様々な分野において扱われている．例えば，2 標本検定 [2]〜[4] は，2 つのデータ集合が異なるデータ集合であるかどうかを検定する問題であり，統計学において長い研究の歴史がある [5]．データマイニングのコミュニティにおいては，コンセプトドリフト分析 [6]〜[8] の文脈で多くの研究が行われている．コンセプトドリフトとは，基本的に，時間により変化するラベル付きの訓練データ集合を前提に，教師付き学習器の変化を考察するものである．しかしながら，大抵の従来手法は，変化の有無を導くことに着目している．それ自体は貴重な情

† 日本アイ・ビー・エム (株) グローバル・ビジネス・サービス事業, matuzawa@jp.ibm.com
IBM Japan, Global Business Services
†† 日本アイ・ビー・エム (株) 東京基礎研究所, {hido,goodidea}@jp.ibm.com
IBM Research - Tokyo
††† 東京大学大学院 情報理工学系研究科, kashima@mist.i.u-tokyo.ac.jp
Graduate School of Information Science and Technology, The University of Tokyo

報を与えるが，先に述べたように，実用上重要なのはむしろ，変化が起こったとして，それはいかなる変化だったのかを詳細に説明することである．教師無し学習の枠内で問題を捉えている限り，この問いに答えるのは簡単ではない．

本論文の構成は以下の通りである．2. 節において，変化解析問題を改めて定義し，我々のアプローチの概要を説明する．意外なことに，3. 節で詳細を述べるが，この教師無し学習の問題は教師付き学習器を用いて解くことができる．提案手法の有効性を確認するため，4. 節にて，実データを用いた実験結果を示す．最後に，5. 節にて関連研究を概観し，6. 節にて本論文を要約する．

2. 問題定義と概要

本章では，変化解析の定式化を行い，我々のアプローチの概要を説明する．

2.1 変化解析問題

ラベル無しのデータからなる 2 つのデータ集合，$\mathrm{X_A} \equiv \{\boldsymbol{x}_{\mathrm{A}}^{(1)}, \boldsymbol{x}_{\mathrm{A}}^{(2)}, \ldots, \boldsymbol{x}_{\mathrm{A}}^{(N_{\mathrm{A}})}\}$ 及び，$\mathrm{X_B} \equiv \{\boldsymbol{x}_{\mathrm{B}}^{(1)}, \boldsymbol{x}_{\mathrm{B}}^{(2)}, \ldots, \boldsymbol{x}_{\mathrm{B}}^{(N_{\mathrm{B}})}\}$，が与えられたとする．ここで，$N_{\mathrm{A}}$ と N_{B} は，それぞれ $\mathrm{X_A}$ と $\mathrm{X_B}$ に含まれるデータ数を表している．

$\boldsymbol{x}_{\mathrm{A}}^{(i)}$ と $\boldsymbol{x}_{\mathrm{B}}^{(i)}$ のそれぞれは，d 次元の特徴空間において同一分布に従う互いに独立 (i.i.d.) なサンプルである．

本論文において，これらのデータ集合に対して，2 つの問題を取り上げる．最初の問題は，2 標本検定と基本的に同じ変化検出問題である．

［Definition 1］（変化検出問題） 同一でないデータ集合 $\mathrm{X_A}$ と $\mathrm{X_B}$ が与えられた時，その差分が大きいかどうかを判定し，$\mathrm{X_A}$ と $\mathrm{X_B}$ の不一致度を計算すること．

ここで，この問題では，コンセプトドリフトとは違って，ラベル無しのデータに焦点を合わせていることに注意して欲しい．

2 番目の問題は，変化解析問題である．

［Definition 2］（変化解析問題） 同一でないデータ集合 $\mathrm{X_A}$ と $\mathrm{X_B}$ が与えられた時，個々の変数に関して差異を説明するような決定ルールの集合を出力すること(注1)．

決定ルールを取得するための教師となる情報が何も与えられていないので，これらは教師無し学習のタスクとなる．

これらの 2 つの問題の違いを理解するため，変化検出問題に対する標準的なアプローチと考えられている 2 標本検定の限界について考えてみる．2 標本検定は，2 つのデータ集合の差分を検出するための統計的な検定方法である．形式的には，P_{A} と P_{B} をそれぞれ $\mathrm{X_A}$ と $\mathrm{X_B}$ に対する確率分布とする時，$P_{\mathrm{A}} = P_{\mathrm{B}}$ であるか，$P_{\mathrm{A}} \neq P_{\mathrm{B}}$ であるかを判定しようとするものである．

統計学分野では，2 標本問題は 2 つのカテゴリーに分類される [5]．最初のカテゴリーはパラメトリック法であり，確率密度分布をモデル化するためにパラメトリックな関数形が明確に仮定される．しかしながら，現実世界のデータの分布は単純な関数形を持たないのが通例であり，現実には，そのような密度のモデリングは一般に困難である．加えて，ガウス分布などの適切なパラメトリックモデルが得られたとしても，変数が互いに独立でないならば，個々の変数に関して差異の発生原因を説明することは，一般には難しい．

2 つ目のカテゴリーは，ノンパラメトリック法であり，密度モデル無しに統計的検証を行うものである．もし，我々の興味が 2 つのデータ集合間の一致の度合いを知ること，すなわち 2 標本検定であるならば，最大平均差異 (maximum mean discrepancy) [4]，Kolmogorov-Smirnov 統計量 [2]，ある種のエネルギー指標による方法 [9]，最近傍統計量 [3]，などを用いて，不一致度を計算することが可能であり，それにより，変化検出問題を解くことができるはずである．しかしながら，これらの方法で変化解析問題を解くことは一般には難しい．なぜなら，ノンパラメトリックなアプローチは密度モデリングを避けるため，データの内部構造に関して得られる情報がほとんどないからである．いくつかの 2 標本検定では標本数の大きさが無限大の極限において漸近分布を得ることができるものの，実用上の変化解析問題を解くことは非常に困難である．何故ならば，そのような分布は漸近分布であるため，標本数が有限である現実世界のデータに対しては適切なモデルには一般にはなり得ないためである．

2.2 概要と我々のアプローチ

2 標本検定の限界を考え，我々はこれらの 2 つの問題に対して，簡単なアプローチを提案した [1]．我々のキーとなるアイディアは次の通りである．$\mathrm{X_A}$ の各データに $+1$ という仮ラベルを付与し，$\mathrm{X_B}$ の各データに

(注1)：オンライン問題に制限する必要がないので，ここでは，“差分解析” という用語がより適切と思われる．しかしながら，(良く知られた専門用語である）変化検出との対比を強調するため，我々はコンセプト変化解析と呼ぶ．

-1 という仮ラベルを付与する．そして，教師付きの方法を用いた分類器を用いて訓練を行う．以後，この分類器を仮想分類器 (VC: virtual classifier) と呼ぶ．

図 1 に我々のアプローチの概要を示す．図中の ◯ と □ は，それぞれ $+1$ と -1 を表す．我々のアプローチにおいて，もし 2 つのデータ集合に実際に差異があるのであれば，それらは分類器により正しく分類されるはずである．従って，高い分類精度 p は，$\mathrm{X_A}$ と $\mathrm{X_B}$ の間の差異を表す．例えば，もし，$P_\mathrm{A} = P_\mathrm{B}$ であるとすると，$N_\mathrm{A} = N_\mathrm{B}$ である時，分類精度はおよそ 0.5 になるはずである．しかしながら，もし p が 0.5 よりも非常に大きいのであれば，我々が推論できることは，ラベルが異なっており $P_\mathrm{A} \neq P_\mathrm{B}$ であるということである

さらに，変化解析問題を解くため，我々は分類器の説明能力の高さを利用する．例えば，ロジスティック回帰のアルゴリズムは，各点の重要性を重みとして与えている．また，別の例では，決定木を用いた場合には，根に近いノードに現れる変数が，分類のための重要の要素となっている．このように，仮想分類器から変化に関する決定規則を得ることができる．

この仮想分類器の利点は，次の通りである． **第一に，変化検出と変化解析の両方を一つの変化解析アルゴリズムの中で実施できる．** 分類器は，分類精度を用いて変化の大きさを容易に与え，属性選択機能を通して変化に対する診断情報を提供することができる．第二に，仮想分類器のアプローチは密度の見積もりが不要である．密度を見積もることは，高次元データに対しては，非常に困難である．最後に，仮想分類器のアプローチは，2 項検定により変化の大きさを評価することができる． **伝統的なノンパラメトリックな 2 標本検定は，漸近分布 (サンプル数が無限大の極限で漸近的に厳密になる分布) を前提にしているため，サンプル数がそれほど多くない実データに対しては，2 標本検定よりも本手法の方が，より有益である．**

3. 仮想的な分類器による変化解析

この章では，我々の提案する教師付き学習を用いた変化解析法の詳細について述べる．上線のついた記号を，仮想的なラベルのついたデータ集合のことであるとする．たとえば，$\bar{\mathrm{X}}_\mathrm{A} \equiv \{(\boldsymbol{x}_\mathrm{A}^{(i)}, +1) \mid i = 1, ..., N_\mathrm{A}\}$, $\bar{\mathrm{X}}_\mathrm{B} \equiv \{(\boldsymbol{x}_\mathrm{B}^{(i)}, -1) \mid i = 1, ..., N_\mathrm{B}\}$. である．また，仮想分類器による予測精度を p とする．

3.1 変化の検出

2 つのデータ集合を組み合わせた集合 $\bar{\mathrm{X}} \equiv \bar{\mathrm{X}}_\mathrm{A} \cup \bar{\mathrm{X}}_\mathrm{B}$ と，二値分類学習のアルゴリズム L が与えられているとする．L を用いて，データ集合 $\bar{\mathrm{X}}$ に対して学習を行い，k-交差検定により分類精度 p を測る．特に，$\bar{\mathrm{X}}$ を k 等分し，そのうち一つをテスト用に，残りの $(k-1)$ 個を学習用に用いる．テスト用に用いるデータ集合を替えながら学習を行い，得られた k 個それぞれのテストにおける分類精度の平均を p とする．

もしも $P_\mathrm{A} = P_\mathrm{B}$ の場合には，L によって得られた分類器による分類は，$\bar{\mathrm{X}}$ におけるベルヌーイ試行として見ることが出来る．従って，$\bar{\mathrm{X}}$ に対する $N_\mathrm{A} + N_\mathrm{B}$ 回の試行の対数尤度の和は，

$$\ln \left[q^{N_\mathrm{A}} (1-q)^{N_\mathrm{B}} \right]$$

となる．ここで，q はクラスラベルが A である確率であるとする．これを q について微分して 0 と置くことによって，最尤解 $q = N_\mathrm{A}/(N_\mathrm{A} + N_\mathrm{B})$ が得られる．分類の精度 p は $\max\{q, 1-q\}$ であるから，p は以下の p_bin によって与えられるはずであることが分かる．

$$p_\mathrm{bin} \equiv \frac{\max\{N_\mathrm{A}, N_\mathrm{B}\}}{N_\mathrm{A} + N_\mathrm{B}} \tag{1}$$

一方，$P_\mathrm{A} \neq P_\mathrm{B}$ の場合，仮想的なクラスラベルが情報量を持っていたということを意味するので，分類精度は p_bin よりも有意に高くなるはずである．特に，2 つのデータ集合の違いが大きいほど，分類精度も高くなる．帰無仮説が，分類精度が p_bin であるとすることによって，二項検定によってこれを検定できるところが，提案手法の大きな特徴の一つである．なお，一般性を失うことなく，$N_\mathrm{A} > N_\mathrm{B}$ であるものとする．ある信頼水準 $\alpha > 0$ に対して，

$$\sum_{n_\mathrm{A}=Np}^{N} \frac{N!}{n_\mathrm{A}!(N-n_\mathrm{A})!} p_\mathrm{bin}^{n_\mathrm{A}} (1-p_\mathrm{bin})^{N-n_\mathrm{A}} \leqq \alpha, \tag{2}$$

であるときに帰無仮説を棄却する．なお，$N = N_\mathrm{A} + N_\mathrm{B}$ であるとする．帰無仮説が棄却された場合，p が p_bin よりも有意に大きかったということであり，これは仮想的なクラスラベルが十分な情報量を持っていた，つまり 2 つのデータ集合には十分な違いがあったということを意味する．棄却するかどうかを判断する p の閾値を，正の定数 γ を用いて $(1+\gamma_\alpha)p_\mathrm{bin}$ のように表現すると，判定条件は以下のように求まる．

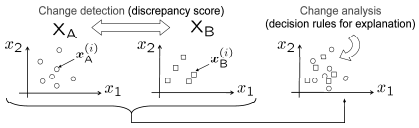


図 1　仮想分類器を用いた我々のアプローチの概要
Fig. 1　High-level overview of the virtual classifier approach.

$$p < (1 + \gamma_\alpha) p_{\rm bin}. \tag{3}$$

例えば，$N = 1000$ で $p_{\rm bin} = 0.5$ の場合，5%有意水準と 1%有意水準はそれぞれ，$\gamma_{0.05} = 0.054$ と $\gamma_{0.01} = 0.076$ になる．十分に大きい N に対しては γ_α の計算に正規分布による近似を用いることが出来る [5].

3.2　変化解析アルゴリズム

二項検定によって 2 つのデータ集合 X_A と X_B の違いが検出された場合，$\bar{\mathrm{X}}$ 中の全てのデータを用い，学習アルゴリズム L（もしくはその他のアルゴリズム）によって再学習を行う．もし，属性のいずれかが分類器において重要な役割を担っていた場合，これらが 2 つのデータ集合の違いを説明する鍵となるはずである．例えば，L として決定木のアルゴリズムである C4.5 [10] を用いるとする．C4.5 は情報量利得の大きい属性を使って，データ集合を再帰的に分割していくアルゴリズムである．得られるモデルは木構造を有しており，根に近い属性ほど重要な属性であるといえる．このように，L が属性選択もしくは重み付けの機能を持つ場合，その能力を変化解析に利用することができる．

図 2 は，我々の提案する変化解析のアルゴリズムをまとめたものである．アルゴリズムの前半部分（1〜3 行）は，二項検定によって変化検出を行う部分であり，後半部分（4 行と 5 行）は，変化解析の部分である．入力パラメータは，信頼水準 α と交差検定の分割数 k の 2 つである．

3.3　クラスラベル付きデータへの適用

図 2 のアルゴリズムはクラスラベルなしのデータに対するものであるが，これをラベル付きデータに対しても適用できるよう拡張することが出来る．この拡張は,「分類器の変化」を解析することが可能になるという点で非常に重要である．分類アルゴリズム M と，以下によって定義される 2 つのラベル付きデータ集合 $\mathrm{D_A D_B}$ が与えられているものとする．$\{(\boldsymbol{x}_{\rm A}^{(i)}, y_{\rm A}^{(i)}) | i = 1, ..., N_{\rm A}\}$ $\{(\boldsymbol{x}_{\rm B}^{(i)}, y_{\rm B}^{(i)}) | i = 1, ..., N_{\rm B}\}$ ここで，$y_{\rm A}^{(i)}$ や $y_{\rm B}^{(i)}$ などはクラスラベルを表すものとする．M によって，$\mathrm{D_A}$ と $\mathrm{D_B}$ から，それぞれ分類器 $\mathrm{M_A}$ と $\mathrm{M_B}$ を得る．ここで行いたいのは，$\mathrm{M_A}$ と $\mathrm{M_B}$ の変化解析，つまり，2 つの分類器 $\mathrm{M_A}$ と $\mathrm{M_B}$ の違いを説明することである．

この問題を解決するために，次の方法によってデータ集合を作成する．各データ $\boldsymbol{x}_{\rm A}^{(i)}$ あるいは $\boldsymbol{x}_{\rm B}^{(i)}$ に対して，2 つの分類器 $\mathrm{M_A}$ と $\mathrm{M_B}$ による分類を行う．もしも 2 つの分類器の予測が食い違えば，このデータを $\mathrm{X_A}$ に加え，そうでなければ $\mathrm{X_B}$ に加える．全てのデータに対してこれを行うことによって，2 つのラベルなしデータ集合が得られることになる．$\mathrm{X_A}$ は 2 つのモデル $\mathrm{M_A}$ と $\mathrm{M_B}$ の食い違いを，$\mathrm{X_B}$ は一致部分を表しているデータ集合であり，これらに対して図 2 の変化解析アルゴリズムを適用することで，分類器の差についての解析を行うことができる．

$$\rho \equiv N_{\rm inc}/(N_{\rm A} + N_{\rm B}) \tag{4}$$

は，$\mathrm{M_A}$ と $\mathrm{M_B}$ の（もしくは $\mathrm{D_A}$ と $\mathrm{D_B}$ の）食い違いの度合いを表す量である．ここで $N_{\rm inc}$ は，予測が食い違うデータの数であるとする．

もともとのクラスラベル付きデータ集合において，クラスラベルの数が多くない場合には，提案手法をさらに複数クラスの分類に拡張することも可能である．例として，もともとのクラスラベルが二値 $(y_{\rm A}^{(i)}, y_{\rm B}^{(i)} \in \{\pm 1\})$ である場合を考える．予測が食い違うデータ集合 $\mathrm{X_A}$ を，さらに 2 つのデータ集合 $\mathrm{X_{A1}}$ と $\mathrm{X_{A2}}$ に分ける．ここで，$\mathrm{X_{A1}}$ は一方の分類器がクラスラベル $+1$ を予測し，もう片方が -1 を予測したデータからなるデータ集合，一方，$\mathrm{X_{A2}}$ は前者の分類器が -1 を，後者が $+1$ を予測するようなデータからなるデータ集合とする．こうして得られた 3 つのデータ集合 $\mathrm{X_{A1}}$，$\mathrm{X_{A2}}$ および $\mathrm{X_B}$ に対して，3 クラスの分類学習アルゴリズム L を適用して，これらを分類する分類器を得たのち，これを解析するとことによって，変化の種類を考慮した，より詳細な解析を行うことができる．

アルゴリズム: 変化解析アルゴリズム
入力:
· 2 つのデータ集合 $\mathrm{X_A}$ および $\mathrm{X_B}$
· 二値分類学習アルゴリズム L
· 交差検定の分割数 k
· 信頼水準 $\alpha > 0$
1. $\mathrm{X_A}$ に属するデータに正例ラベル A を，$\mathrm{X_B}$ に属するデータに負例ラベル B を与える．
2. 学習アルゴリズム L を用いた k-交差検定によって，分類精度 p を得る．
3. $p < p_{\mathrm{bin}}(1 + \gamma_\alpha)$ の場合，アルゴリズムを終了する．
そうでない場合には，$\mathrm{X_A}$ と $\mathrm{X_B}$ に十分な差があると結論する．
4. 全てのデータを用いて L を再学習する．
5. 得られたモデルを精査することによって，$\mathrm{X_A}$ と $\mathrm{X_B}$ の差についての知見を得る．

図 2　仮想分類器を用いた変化解析アルゴリズム
Fig. 2　The virtual classifier algorithm for change analysis

4. 実　　験

我々は，人工データ，及び，実データを用いて，変化解析に対して我々の仮想分類器を用いたアプローチの有用性評価を行なった．以下の実験において，特に言及がなければ，$\alpha = 0.05$ 及び $k = 10$ を用いる．分類アルゴリズム L として，我々は，主に Weka [10] の `J48` として実装された C4.5 の決定木作成アルゴリズムを適用した．このツールは，各リーフについて最小インスタンス数を表す $minNumObj$ と呼ばれるパラメータを持っている．2 標本間の線形分離性の度合いを見るために，さらに，Weka で `Logistic` として実装されているロジスティック回帰 (LR: Logistic Regression) を用いた．`Logistic` における 2 つのパラメータ (ridge パラメーター，最大繰返数) は，それぞれデフォルト値 (10^{-8}，及び，無限大) を用いた．

4.1 人工データ

はじめに，我々は変化解析が容易なデータを用意し，提案手法が変化をきちんと解析できることを示すために，最初の実験を行なった．我々は，$N_{\mathrm{A}} = N_{\mathrm{B}} = 500$ であるような人工データを用いて 2 つの実験を行なった．このデータサイズに対して，臨界精度を 0.527 ($\gamma_{0.05} = 0.054$) と与えた．どちらの実験においても 10 個の属性はガウス分布に従って付与され，それぞれが独立である．

最初の実験において，データ集合 $\mathrm{X_A}$ と $\mathrm{X_B}$ は，P_{A} と P_{B} が大きく異なるように生成した．$\mathrm{X_A}$ は，$Attr1$（最初の属性）を除き，属性は全て標準偏差が 1.0 となるようにし，$Attr1$ は 4.0 になるようにした．一方，$\mathrm{X_B}$ は，$Attr2$（2 番目の属性）を除く全ての属性の標準偏差を 1.0 とし，$Attr2$ の標準偏差を 4.0 になるようにした．図 3 (a) に属性 $Attr1$ と属性 $Attr2$ の 2 次元空間におけるデータ集合の周辺分布を示した．我々の目的は，二つのデータ集合の差異について，その原因となる属性として $Attr1$ と $Attr2$ を取り出すことにある．

このデータ集合に対して $minNumObj = 10$ とする決定木を作成し，変化解析を行なった．生成された決定木の予測精度 p は，10 分割交差検定による計算では 0.797 であり，臨界精度を遥かに超えている．これは，2 つのデータ集合が異なっていることが正しく判断されたことを示している．図 3 (b) は，仮想分類器としての決定木を示している．図において，楕円とエッジのラベルは，それぞれ分割された属性と決定ルールを表している．影付きの四角は，クラスラベルと (1) ノードに入ったインスタンスの数と (2) ノードに間違って入ったインスタンスの数を (1)/(2) 形式で表している．ただし，(2) が 0 個の場合は，省略される．Fig. 3 (a) 内の線は，決定木によって分割された境界を示している．明らかに，$Attr1$ と $Attr2$ の間の意図的な非線形な変化を決定木が学習していることがわかる．注意すべきは，L として LR を用いた時，10 分割交差検定による精度 p は 0.505 であり，これは臨界精度を下回っている．この結果は，非線形な決定境界の役割が重要であることを明らかに示している．

2 つ目の実験では，P_{A} と P_{B} が同じになるようにした．$\mathrm{X_A}$ と $\mathrm{X_B}$ の両方において，属性 $Attr2$ を除く全ての属性の標準偏差を 1.0 とし，属性 $Attr2$ の標準偏差を 4.0 とした．図 4 は，図 3 (a) に相当する周辺分布を示している．最初の実験とは対照的に，決定木のモデルは 0.5 という低い予測精度となっており，これは，統計的に大きな差分が存在していないことを示している．この結果は，決定木を用いた我々のアプローチが，データ集合がクラス間の差分を含む時だけ，統

計的に重要な差分を示すという妥当な分類器を生成したことを示している．

4.2 エンロン社メールデータ

我々が用いたエンロン社メールデータ集合とは，米国の倒産したエンロン社 (Enron Corporation) における実際のメールのアーカイブであり，クラスラベルは付与されていない [11]．我々が用いた 2001 年度のメールデータ集合には，272,823 件のメールメッセージを含んでおり，そのコンテンツは単に単語の集合 (bag-of-words) として与えられている [12]．我々は，データ集合を 2001 年の上半期 (1H) と下半期 (2H) に分割し，各半期における 100 個と 150 個の最頻出単語からなる特徴ベクトルを生成した．選択した特徴語を含まないゼロベクトル (データ) は，意味がないので削除した．各半期のデータは，四半期の比較ができるように，さらに，2 つに分割した．我々は，$numMinObj = 1,000$ として上半期と下半期のそれぞれで変化解析を実行した．例えば，下半期 (2H) の分析において，X_A と X_B は第 3 四半期 (3Q) と第 4 四半期 (4Q) に該当する．

表 1 に，見積もった予測精度を示す．ロジスティック回帰 (LR) と決定木 (DT) の精度がどちらも臨界精度を遥かに超えていることがわかる．差分の詳細を調査するため，我々は下半期 (2H) データに着目し，変化解析を行い，図 5 の決定木を得た．出力された決定木はルートから上位 5 つのノードを選択し，100 単語のモデルと 150 単語のモデルを比較した．決定木の記法は，4.1 章と同じである．各属性のランクはここに追記した (‘access’ が 44 番目に高頻度であったなど)．閾値は，各メール中の特徴語の頻度を表している．我々は，単に頻度を基本とする属性の生成を行なう戦略をとったので，150 単語の決定木が特定の意味を生み出す単語を含む傾向にあった．

‘position’ は，その頻度が 144 位であるにも関わらず，150 単語モデルにおける根ノードに出現している．エンロン社は 2001 年末に倒産した．もし，役職を失う運命にある社員によってどのような会話が成されたのかを想像できれば，その結果は，とても示唆に富んでいる．さらに，‘Jeff’ と ‘Davis’ は，第 4 四半期を最も特徴付ける属性であった．興味深いことに，当時，エンロン社の CEO の名前は Jeffrey Skilling であり，彼は，突然，2001 年の 8 月にその職位を辞しており，それは全てのストックオプションを売却した後であった．多くの社員が，その時に彼に関する色々な会話をしたに違いない．Davis に関してであるが，当時，Gray Davis という重要人物がおり，同じ年にカリフォルニアの電力危機の渦中にあったカリフォルニア州知事である．これは，第 4 四半期のエンロン社の投資に対する責任を問われた結果であろう．仮想分類器が新聞の情報などに無しに，これらの重要人物を発見し，エンロン社のような複雑系の変遷を学習する有用性を示したことに注意して欲しい．

4.3 学会活動データ

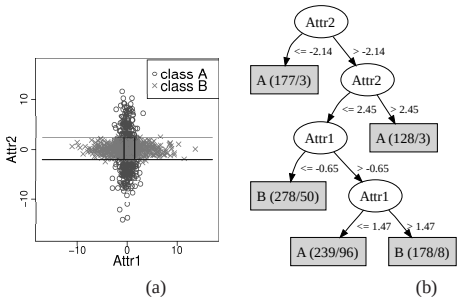
ラベル付きのカテゴリカルデータに対するアプリケーションの例として，ある企業の研究部門で収集された “学会活動” データに対する変化解析を行なった．このデータ集合は，5 年分の 4,683 件の $(\boldsymbol{x}^{(s)}, y^{(s)})$ という形式のデータからなり，s は時間軸であり，$y^{(s)}$ は’Y’（重要）か’N’（重要でない）のどちらかのラベルを表している．各ベクトル $\boldsymbol{x}^{(s)}$ は，*title*，*group* 及び，*category* という 3 つのカテゴリカル属性を含んでおり，その値については，表 2 に示した．

’Y’ または’N’ のラベルは，各 $\boldsymbol{x}^{(s)}$ で与えられる各活動を評価して人手により付与されており，入力時の学会活動データベース管理者の主観的な意思に依存する．例えば，何人かの管理者はアジア地区での国際会議を重要と判断するかも知れないが，他の管理者はそう判断しないかも知れない．学会活動データベース管理者の交代やデータベースへの入力ガイドラインの変更などのイベントの発生により，’Y’ または’N’ のラベルの判断基準が変更されてしまっていた．この分析の目的は，重要ラベル (’Y’) を選択する基準の変更が何時，どのような変化が起きたかを調査することにある．

我々は，データを四半期毎に分割して，$D_1, D_2, \ldots, D_{14}$ で表す 14 個のサブセットを作成した．最初に，明確なコンセプトドリフトが発生しているかどうかを見るため，隣接する四半期との間で式 (4) に示す不一致スコア ρ を計算した．具体的には，$t = 1, 2, ..., 13$ に対する D_t と D_{t+1} を D_A と D_B として計算した．

M として，決定木を適用した．図 6 に，全ての組み合わせに対する不一致度スコア ρ を示す．2 つのピークが $t = 5$ と $t = 10$ の周辺に現れたのを見ることができ，これは，これらの期間の間にコンセプトドリフトが明らかに発生していたことを示している．興味深いことに，これらのピークは学会活動データベースの管理者が実際に交代した時期と重なっており，引継ぎがうまく行かなかったことを示唆している．

次に，$t = 5$ の前後で何が起きたのかを調べるため，D_5 と D_6 を用いて，変化解析を行なった．3.3 章の手



(a) (b)

図 3　(a) $Attr1$ と $Attr2$ の分布，(b) 仮想分類器による結果
Fig. 3　(a) Distribution over $Attr1$ and $Attr2$ in the first synthetic data, and (b) the resulting virtual classifier.

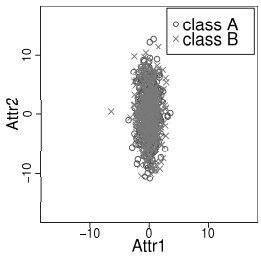


図 4　$Attr1$ と $Attr2$ の分布
Fig. 4　Distribution over $Attr1$ and $Attr2$ in the second synthetic data.

表 1　エンロンデータに対する予測精度
Table 1　Prediction accuracies on Enron

Data set		アルゴリズム	
期間	単語	LR	DT
2001-1H	100	64.3%	67.4%
2001-1H	150	65.4%	68.4%
2001-2H	100	60.9%	62.8%
2001-2H	150	62.3%	64.1%

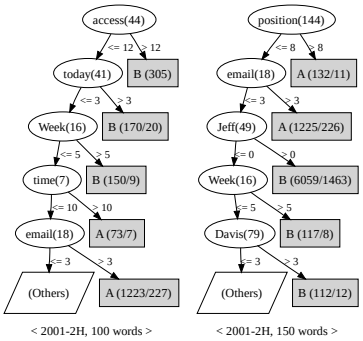


図 5　仮想分類器（エンロンデータ 2001-2H)
Fig. 5　VCs on the Enron 2001-2H data set

表 2　学会活動データにおける 3 つの属性とその値
Table 2　Three features and their values in the academic activity data

category	*title*	*group*
GOVERNANCE, EDITOR, ORGANIZATION, PROFESSIONAL ACTIVITY	COMMITTEE, MEMBER, AWARD, OTHERS	UNIVERSITY, DOMESTIC, STANDARD, PUBLISHER, SOCIETY1, OTHERGROUPS

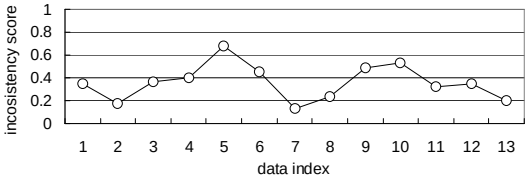


図 6　D_t と D_{t+1} 間の不一致度スコア ρ.
Fig. 6　Inconsistency scores ρ between D_t and D_{t+1}.

順に従い，図 7 に示した仮想分類器を得た．ここで生成した決定木は 3 つのクラスから構築されている．3 つのクラスは，全データのサブセット X_{A1}，X_{A2}，X_B であり，X_{A1} は，予測値が Y→N というラベルの変化があったものである．また，X_{A2} は，N → Y という変化があったデータの集合，X_B は，Y → Y とラベルに変化が無かった集合である．図 7 において N から Y に変化した ‘NY’ というリーフを見ることで，D_5 と D_6 の間で起きた興味深い変化を見つけるができる．例えば，学会活動データベースの $t = 6$ における新しい管理者は，ジャーナルの編集者と同じようにプログラム委員や実行委員などの活動を重要と考え’Y’ と入

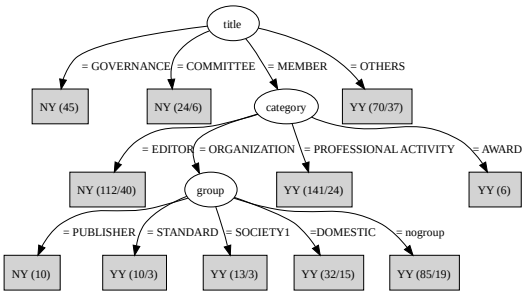


図 7　D_5 と D_6 に対する仮想分類器
Fig. 7　Virtual classifier for D_5 and D_6.

力する傾向にあったことが分かる．

D_5 と D_6 からそれぞれ作成した 2 つの決定木 M_5 と M_6 を直接比較すれば良いと考えることができるかも知れない．しかしながら，決定木の複雑な木構造を考えると，異なる決定木を直接比較することは一般に困難である．我々の仮想分類器は，異なる分類器間の差異を見る直接的な手段を与えており，直接比較するというようなナイーブなアプローチとは対照的である．

5. 関 連 研 究

教師付き分類学習と変化点検出の関係は，二標本問題への最近傍分類法の適用という文脈で 1980 年台に暗に示されている [3] ものの，変化の内容を調べるのが目的である変化解析については触れられていない．実際，最近傍分類器は，明示的なモデルを構成しないため，変化解析には向いていない．

2 つのデータ集合の差を測るための別の枠組みとしては FOCUS [13] がある．教師付き学習の場合，FOCUS は 2 つのデータ集合それぞれに対して決定木を構成し，これらが同一の構造を有するようになるまでこれらを伸長する．そして，決定木の同じ葉に落ちるデータの個数によって，データ集合の違いの度合いを測る．しかしながら，高次元の場合には，伸長によって作られた木のサイズは指数的に大きくなりうるため，膨大な計算資源を必要とするという欠点がある．

原因分析の文脈では，しばしばベイジアンネットワーク [14] などのグラフィカルモデルが用いられる．ベイジアンネットワークにおいても，2 つのデータ集合のどちらであるかを示す変数を加えることによって，原理的には変化解析を行うことが可能である．しかしながら，一般的にグラフィカルモデルの学習は，すべての変数の同時分布の推定を行っており，その構造推定には，多くのデータと計算時間が必要とする．一方，我々の手法は，データ集合を示す変数を直接的に説明するモデルを構成するため，遥かに効率がよいという利点がある．

ストリームデータの解析において，コンセプトドリフトの扱いは本質的な課題であり，多くの研究が行われてきた [6]〜[8]．最近では，Dries らが我々のアプローチを元に,1-norm Support Vector Machine を分類器として用いたコンセプトドリフト検出手法を提案した [18]．**変化解析についての研究は [19] [15] などの研究が行われている．Song [19] らは，異なる時期のデータから作成された相関ルールを取得することで，特定の顧客層 (プロファイル) に対する購買行動の変化を分析する手法を提案している．この手法は，特定の顧客層の変化解析は行うには有益であると考えられるが，顧客行動全体の変化を解析することには適していない．**

KBS-stream [15] はコンセプトドリフトの程度を測るだけでなく，その違い説明する差分モデルと呼ばれるモデルを与える．**KBS-Stream では，複数のモデルの重み付けの変更にすることにより学習を行なっており，現在のモデルが予測を誤ったデータ集合において，正例と負例を正しく判別するように新しいモデルが構築されるため，差分モデルは，複数のベースモデルの重みがどのように変化したかを示すモデルとして与えられる．一方，我々の手法においては，現在のモデルが正しい予測を行なったデータ集合と，予測を誤ったデータ集合が判別されるように構成され，その差分が差分データとして与えられる．KBS-Stream は，ストリームデータに対するコンセプトドリフトを見るアルゴリズムとしては，変化の大きさを検出する点やパフォーマンスなどの点で優れているが，差分モデル**

を構成するどの属性が具体的に変化に寄与しているのかを解析する点においては，我々の手法が優れていると考えられる． どちらのモデルも異なった視点から，コンセプトドリフトの解析にアプローチしているといえる．

また，教師付き学習と教師無し学習を組み合わせたものとして Yamanishi [20] らの研究がある．Yamanishi らは，教師無し学習により外れ値の検出を行い，検出された外れ値をラベルとして与えることにより，教師付き学習を行った外れ値検出ルールの抽出を行っている．アプローチは非常に近いが，本論文の目的とは異なっている．

他にも，類似の問題に取り組んでいる研究としては，モデルのアンサンブル平均 [16] や高速な決定木 [17] を用いたものなどがあるものの，これらは変化解析とは本質的に異なっている．

6. おわりに

本論文で，我々は，新しいデータマイニングの問題である変化解析問題と，その解法を提案した．提案手法は，もし，2 つのデータ集合が別々の分布から得られたサンプル集合であるならば，2 つのデータ集合それぞれに属するデータに仮想的に正例と負例のクラスラベルを振った場合に，これらを高い精度で分類する分類器を作ることができるはずであるというアイディアに基づいている．我々は，データ集合の違いの度合いは，学習したモデルを用いた二項検定によって検定することができることを示した．また，得られた分類器は，2 つのデータ集合の違いを説明するモデルになっているため，モデルを調べることによって，2 つのデータ集合の違いについての知見を得ることが出来ることを示した．また，人工データ集合と実データ集合を用いた実験結果は，提案手法を用いることによって，データ集合の違いについての興味深い知見が得られることを示している．

我々が実験で用いたデータに対して，変化検出，及び，変化解析を行うことが出来た．しかしながら，与えられた問題の種類や仮想分類器の性能によっては，変化検出が困難なケース，変化検出までは可能であっても変化解析が困難なケースなどが存在すると考えられる．実際に，変化検出，変化解析が出来るかどうかを問題の特徴などから見極めることは困難であり，本手法が適用可能な問題の定義は今後の研究課題と考えている．また，一つの仮想分類器が，全ての与えられた問題に対して有効というわけではないので，適切な仮想分類器を選択する必要があると考えている．本論文で用いた仮想分類器以外の選択肢も当然考えられるであろう．コンセントドリフトには様々なケースがあり，各ケースに対して最適な分類器を厳密に定義することは難しいと考えており，最適な分類器を見極めることも今後の課題である． また，ストリームデータにおける変化解析問題や，回帰問題における変化解析問題なども，興味深い研究課題である．

謝　辞

学会活動データの解析において久保晴信博士に多大な協力を得た．ここに謝意を表する．

文　献

[1] Hido, S., Idé, T., Kashima, H., Kubo, H., Matsuzawa, H.: Unsupervised change analysis using supervised learning. In: Proc. the 12th Pacific-Asia Conf. Knowledge Discovery and Data Mining. (2008) 148–159

[2] Friedman, J., Rafsky, L.: Multivariate generalizations of the Wolfowitz and Smirnov two-sample tests. Annals of Statistics **7** (1979) 697–717

[3] Henze, Z.: A multivariate two-sample test based on the number of nearest neighbor type coincidences. Annals of Statistics **16** (1988) 772–783

[4] Gretton, A., Borgwardt, K., Rasch, M., Schölkopf, B., Smola, A.J.: A kernel method for the two-sample-problem. In: Advances in Neural Information Processing Systems 19. MIT Press (2007) 513–520

[5] Stuart, A., Ord, J.K.: Kendall's Advanced Theory of Statistics. Volume 1. Arnold Publishers Inc., 6th edtion (1998)

[6] Fan, W.: Streamminer: A classifier ensemble-based engine to mine concept-drifting data streams. In: Proc. the 30th Intl. Conf. Very Large Data Bases. (2004) 1257–1260

[7] Wang, H., Yin, J., Pei, J., Yu, P.S., Yu, J.X.: Suppressing model overfitting in mining concept-drifting data streams. In: Proc. the 12th ACM SIGKDD Intl. Conf. Knowledge Discovery and Data Mining. (2006) 20–23

[8] Yang, Y., Wu, X., Zhu, X.: Combining proactive and reactive predictions for data streams. In: Proc. the 11th ACM SIGKDD Intl. Conf. Knowledge Discovery and Data Mining. (2005) 710–715

[9] Zech, G., Aslan, B.: A multivariate two-sample test based on the concept of minimum energy. In: Proceedings of Statistical Problems in Particle Physics, Astrophysics, and Cosmology. (2003) 8–11

[10] Witten, I.H., Frank, E.: Data Mining: Practical Machine Learning Tools. Morgan Kaufmann Publishers Inc. (2005)

[11] Klimt, B., Yang, Y.: The Enron corpus: A new dataset for email classification research. In: Proc. the 15th European Conf. Machine Learning. (2004) 217–226

[12] Other forms of the Enron data: URL: `http://www.cs.queensu.ca/~skill/otherforms.html`.

[13] Ganti, V., Gehrke, J.E., Ramakrishnan, R., Loh, W.: A framework for measuring changes in data characteristics. Journal of Computer and System Sciences **64**(3) (2002) 542–578
[14] Pearl, J.: Probabilistic reasoning in intelligent systems: networks of plausible inference. Morgan Kaufmann Publishers Inc. (1988)
[15] Scholz, M., Klinkenberg, R.: Boosting classifiers for drifting concepts. Intelligent Data Analysis Journal **11**(1) (2007) 3–28
[16] Street, W.N., Kim, Y.: A streaming ensemble algorithm (SEA) for large-scale classification. In: Proc. the 7th ACM SIGKDD Intl. Conf. Knowledge Discovery and Data Mining. (2001) 377–382
[17] Hulten, G., Spencer, L., Domingos, P.: Mining time-changing data streams. In: Proc. the 7th ACM SIGKDD Intl. Conf. Knowledge Discovery and Data Mining. (2001) 97–106
[18] Dries, A., Rückert, U.: Adaptive concept drift detection. In: Proc. the 9th SIAM Intl. Conf. Data Mining. (2009) 233–244
[19] Song, H. S., Kim, J. K., Kim, S. H.: Mining the change of customer behavior in an internet shopping mall. Expert Systems with Applications, Volume 21, Issue 3, Elsevier, pp:157-168, (2001)
[20] Yamanishi, K., Takeuchi, J.: Discovering Outlier Filetering Rules From Unlabeled Data—Combininig Supervised Learners with Unsupervised Learneres-. In: Proc. KDD2001, pp:389-394.

Abstract We propose a formulation of a new problem, which we call *change analysis*, and a novel method for solving the problem. In contrast to the existing methods of change (or outlier) detection, the goal of change analysis goes beyond detecting whether or not any changes exist. Its ultimate goal is to find the explanation of the changes. While change analysis falls in the category of unsupervised learning in nature, we propose a novel approach based on *supervised* learning to achieve the goal. The key idea is to use a supervised classifier for interpreting the changes. A classifier should be able to discriminate between the two data sets if they actually come from two different data sources. In other words, we use a hypothetical label to train the supervised learner, and exploit the learner for interpreting the change. Experimental results using real data show the proposed approach is promising in change analysis as well as concept drift analysis.

Key words change analysis, two-sample test, concept drift.