企業の業績発表記事からの業績要因の抽出

酒井 浩之　　増山 繁
sakai@smlab.tutkie.tut.ac.jp,　masuyama@tutkie.tut.ac.jp
豊橋技術科学大学 知識情報工学系

1　はじめに

本研究では，経済新聞記事から企業の業績発表に関する記事を抽出し，業績要因を自動的に抽出する手法を提案する．例えば,「マンションの販売が好調」のような好調な事業が記述されている部分，及び,「公共工事の減少で鉄管などが不振」のような不振な事業が記述されている部分を抽出する．このような情報は，例えば株式投資などで今後投資すべき業種や企業を決定するために必要である．しかし，企業の業績発表は近年では年に 4 回行う企業も多く，その度に新聞やインターネットで内容が配布されるが，その全てに目を通すことは大変な労力を要する．そこで，本手法では，まず，ある企業に関する記事が業績発表であるかどうかを判定し，さらに，その業績要因を自動的に抽出する．

業績要因の抽出方法は，要因が記述されている部分を示す手がかり表現を自動的に獲得することで行うが，少数の手がかり表現を入力することで自動的に多数の手がかり表現を取得する．関連研究として，峠らは，インターネット掲示板から主観的な評価を表している文を抽出するのに，人の主観的評価を表す単語を評価表現として，評価の対象となる単語を評価表現から 2 つの規則にあてはめ自動獲得し，さらに，予め人手で定めた評価文のパターンを用いて文抽出を行っている [4]．那須川らは，好評文脈，不評文脈を分析し，人手で分析して得られた好不評表現の性質や規則を使用することでネット上の掲示版から好評表現，不評表現を取得する手法を提案している [2]．那須川らの手法では，種表現として少数の好評表現，不評表現を人手で与え，その種表現から文書中の好不評文脈を推測し，その中からさらに好評表現，不評表現を取得することを繰り返して，ブートストラップ的に多くの好評表現，不評表現を自動的に抽出している．手がかり表現の自動取得という点や，人手で種表現を与えてブートストラップ的に新たな手がかり表現を獲得する点は本手法も同様なアプローチであるが，本手法では少数の手がかり表現を人手で与えた後は，人手で作成したパターンなどを使用せず統計的情報を使用して自動的に新たな手がかり表現を獲得している点が異なる．

また，経済新聞記事から企業に関する記事を取得する場合に対応する必要がある企業の略称（例えば，全日本空輸の略称は「全日空」）を新聞記事から自動的に獲得する手法もあわせて提案する．

2　企業の業績発表記事の抽出 (前処理)

本論文で提案する手法は企業の業績発表記事から業績要因を抽出する手法であるが，その前処理として新聞記事コーパスから業績発表記事を抽出する．そして，抽出された業績発表記事からその業績要因を抽出する．新聞記事コーパスからの業績発表記事の抽出には Support Vector Machines(SVM)[5] を用いる．

2.1　訓練データの作成

SVM の学習に用いる訓練データの作成について述べる．訓練データは 00 年の日経新聞記事から，表題，および，本文に企業名 (上場企業 3751 社) かその略称が含まれている記事を抽出し，その中から人手で業績発表記事を判別して正例とした．新聞には企業名の正式名称ではなく，その略称で記述される場合があるため，ある企業に関する記事を抽出する場合，企業名の略称にも対応する必要がある．しかし，1 つの企業の正式名称に対して複数の略称が存在する場合もある．よって，その数は膨大になるため，企業の正式名称とその略称の対応は新聞記事コーパスから自動的に獲得した．獲得手法の詳細は 3 章で述べる．その結果，正例として 2920 個の記事を得た．そして，正例と同数の業績発表以外の記事を無作為に選び，負例とした．

2.2　素性選択

SVM における素性選択について述べる．本タスクにおける素性は正例にのみ多く含まれている語 (名詞，動詞，形容詞) とした．そのために，まず，正例に含まれている語に対して以下の式 1 で重み付けを行い，重みが上位半分となる語を抽出する．

$$W(t_i, S_p) = P(t_i, S_p)H(t_i, S_p) \tag{1}$$

ここで，$P(t_i, S_p)$ は正例の文書集合 S_p における語 t_i の出現確率である．また，$H(t_i, S_p)$ は，正例の文書集合 S_p に含まれる各文書における語 t_i の出現確率に基づくエントロピーを表し，エントロピーが高い語ほど正例の文書集合に均一に分布している語であることが分かる．$H(t_i, S_p)$ は次の式 2 で定義される．

$$H(t_i, S_p) = -\sum_{d \in S_p} P(t_i, d) \log_2 P(t_i, d) \tag{2}$$

$$P(t_i, d) = \frac{tf(t_i, d)}{\sum_{d \in S_p} tf(t_i, d)} \tag{3}$$

ここで，$P(t_i, d)$ は文書集合 S_p において，文書 d における語 t_i の出現確率，$tf(t_i, d)$ は，文書 d に語 t_i が出現する数を表す．

式 1 で表した重みでは，一般的な語であれば業績発表記事とは関係のない語でも高い重みが付与される．しかし，そのような語は負例においても高い重みが与えられる可能性が高い．そこで，正例の場合と同様に，負例に属する文書集合 S_n に含まれる語に対しても重み $W(t_i, S_n)$ を求める．そして，ある語 t_i の重み $W(t_i, S_p)$ が重み $W(t_i, S_n)$ の 2 倍より大きければ，その語 t_i を素性として選択する．以上の処理により，一般的な語が素性として選択されることを防ぐ．SVM による学習に用いる素性ベクトルの各要素は，訓練データの各文書における素性として選択された語の出現確率とした．また，実装にあたり，SVM^{light} [1]を使用した．テストデータは 01 年から 05 年の日経新聞記事のうち表題と本文に企業名，もしくは，その略称が含まれている記事とし，その記事集合から 20880 個の業績発表記事を取得した．

3　企業略称対応の獲得

本章では，ある企業名に対応する略称を新聞記事コーパスから自動的に獲得する手法について述べる．

[1] http://svmlight.joachims.org

3.1 企業略称候補の取得

まず，企業名の略称候補を得る．新聞記事では，記事の本文には正式名称が出現するが表題にはその略称が出現している場合が多い．そのため，新聞記事集合において，企業の正式名称が本文に存在する記事の表題から，その企業に対応する略称候補を以下のように抽出する．

Step 1: 企業の正式名称が本文に存在する記事の表題を単語に分割する．ただし，連続して出現する名詞は結合し複合名詞とする．

Step 2: 各語の先頭一文字目が企業の正式名称に含まれている語を取得する．

Step 3: 取得した語から，企業名の正式名称を構成する文字が，同一順序で連続して出現する文字列を取得し，略称候補とする．

上記の処理により，例えば，企業名として「NTT ドコモ」を本文に含む記事の表題「KDD の設備買収、ドコモ、無線基幹網を増強」から，略称候補として「ドコモ」を得る．しかし，これだけでは多くの誤った略称候補が取得されるため，それらを除去する必要がある．

3.2 誤った略称候補の除去

まず，略称候補が一般名詞であれば除去する．ある略称候補が一般名詞であるかの判定は日本語語彙大系 [1] の単語体系を使用した．また，単語体系に登録されている固有名詞も除去する．ただし，企業名，姓名は除去しない．また，複数の正式名称の略称候補になった略称候補を除去する．これは，略称は 1 つの企業名に対応し，複数の正式名称を指すことはないからである．しかし，例えば「日立」のように一般に「日立製作所」のことを表す略称でありながら,「日立メディコ」「日立マクセル」といった企業名の略称候補として取得される場合もあり，単純に複数の正式名称の略称候補になった略称候補を除去するわけにはいかない．そのため，除去すべき略称候補は，複数の正式名称の略称候補でありながら新聞記事集合からの各正式名称に対する略称候補の取得数が同程度である略称候補である．これを判別するために，ある語 t が正式名称の略称候補となる確率に基づくエントロピーを以下の式 4 から求め，ある閾値よりエントロピーが高い略称候補を除去する．

$$H(t) = -\sum_{c \in C(t)} P(t,c) \log_2 P(t,c) \tag{4}$$

$$P(t,c) = \frac{f(t,c)}{\sum_{c \in C(t)} f(t,c)} \tag{5}$$

ただし，

$P(t,c)$**:** 略称候補を取得する文書集合において， ある語 t が正式名称 c の略語候補となる確率

$C(t)$**:** ある語 t が略称候補となった正式名称の集合

$f(t,c)$**:** 略称候補を取得する文書集合において，ある語 t が正式名称 c の略称候補として取得される頻度

閾値は以下の式 6 で求める．

$$T_t = \beta \log_2 |C(t)| \tag{6}$$

β は，0 から 1 までの定数である．また，エントロピーが閾値以下であるにもかかわらず複数の正式名称の略称候補となっている場合は，$f(t,c)$ が最も大きい c を対応している正式名称であるとする．例えば,「日立」の場合は，最も $f(t,c)$ が大きい「日立製作所」を略称候補「日立」に対応する正式名称とする．

3.3 企業略称の取得

上記手法を実装し企業の正式名称に対応する略称を取得した．企業名に対応する略称を獲得するためのソースとなる新聞記事コーパスとして 1990 年から 2005 年までの日経新聞記事を使用した．正式名称として上場企業 3751 社を使用し，閾値の定数 β を 0.75 とした場合，2603 個の略称を獲得した．その精度を測定したところ，90.8% であった．なお，1 つの正式名称に対して複数の略称が対応する場合もあるため，何らかの略称が存在した企業は 1680 社であった．

4 業績要因の抽出

本章では，企業の業績発表に関する記事集合の中から，その業績要因を自動的に抽出する手法について述べる．なお，業績要因は 1 つの文中の複数の文節で構成される．業績要因の抽出は，文中で業績要因の後に出現する表現を手がかり表現とし，それを使用することで行う．例えば，業績要因として好調な事業が示されている記述は「が好調」という手がかり表現の前に出現していることが多く，不振な事業が示されている記述は「が不振」という手がかり表現の前に出現していることが多い．したがって，これらの有効な手がかり表現を取得できれば，業績要因を自動的に抽出することができる．しかしながら，業績要因を抽出するために有効な手がかり表現は数多く，それらを全て人手で抽出することは困難である．そこで，企業業績発表の記事集合中からの手がかり表現の自動獲得を行う．

4.1 手がかり表現の自動獲得

手がかり表現は，我々が以前提案した交通事故事例記事からの事故原因表現の獲得のための手法 [3] に対して，4.5 節で述べる不適切な手がかり表現を除去するための改良を施し，本タスクに適用して取得した．本手法の概要を以下に示す．

Step 1: 少数の手がかり表現を人手で与え，それに係る節を取得する．

Step 2: 取得した節の集合から，その中で共通して頻繁に出現する表現（一つの文節から複数の文節で構成される場合もある.）を抽出し，それが係る節を新たな手がかり表現として獲得する．

Step 3: 獲得した手がかり表現から，それに係る節を取得する．

Step 4: Step 2，3 を，新たな手がかり表現が獲得されなくなる，もしくは，予め定めた回数まで繰り返す．□

Step 1 では，初期の手がかり表現として「が好調」,「が不振」を人手で与えた．

4.2 共通頻出表現の抽出

Step 2 において，手がかり表現から取得された節から，その節に含まれる共通頻出表現を抽出する．ここで，共通頻出表現とは，業績要因が記述されている節において，その業績要因が異なっている場合でも共通して頻繁に出現する表現と定義する．例えば，3 つの業績発表記事における業績要因がそれぞれ「A の売り上げの増加が寄与」「B の売り上げが好調」「C の売り上げが低迷」であった場合，共通頻出表現は「売り上げ」となる．

まず，手がかり表現に係る文節に対して，それに係る文節を追加することで派生する表現を取得する．そして，既に得られている表現に係る文節を次々に追加することで派生する表現を全て取得する．例えば，文書 A に「新店の紳士服の売り上げが好調」という文が存在していたとす

れば，手がかり表現「が好調」に係る文節 (実際には，助詞「が」を除去して「好調」に係る文節)「売り上げ」と，それに文節を追加して「紳士服の売り上げ」「新店の紳士服の売り上げ」という 3 つの表現を取得する．また，文書 B に「主力のカードゲームの売り上げが好調」という文が存在していたとすれば，この文から「売り上げ」「カードゲームの売り上げ」「主力のカードゲームの売り上げ」という 3 つの表現を取得する．そして，文書 A と文書 B からは「売り上げ」が 2 回,「カードゲームの売り上げ」,「主力のカードゲームの売り上げ」,「紳士服の売り上げ」,「新店の紳士服の売り上げ」が 1 回，派生したことになる．ここで，手がかり表現に直接係っている文節を c とおく．上記の例では「売り上げ」が c となる．

次に，文節 c から派生した各表現 e に対して，以下の式 7 で表されるスコアを計算する．

$$Score(e,c) \quad = \quad -f_e(e,c)\sqrt{f_p(e)}\log_2 P(e,c) \qquad (7)$$

ただし，$f_p(e)$ は表現 e に含まれる文節の数，$P(e,c)$ は文節 c から派生する表現 e の派生確率，$f_e(e,c)$ は文節 c から派生する表現 e の派生回数である．例えば，前述の文書 A と文書 B の例では,「売り上げ」の $f_e(e,c)$ の値は 2 であり，c から派生する表現の総数は 6 であるため，$P(e,c)$ の値は 2/6 となる．そして，c から派生する表現の中で，$f_e(e,c)$ の値が 2 以上である表現のうちスコアが最大の表現を共通頻出表現として抽出する．

4.3 共通頻出表現の選別

ある手がかり表現から共通頻出表現を抽出しても，中には不適切な表現も抽出される．そこで，手がかり表現から抽出された共通頻出表現の中から適切な共通頻出表現を選別する．具体的には，様々な手がかり表現に係っている共通頻出表現は適切であるという仮定に基づき，共通頻出表現が手がかり表現に係る確率に基づくエントロピーを求め，その値が閾値 T_e 以上の共通頻出表現を選別する．共通頻出表現が手がかり表現に係る確率に基づくエントロピーは式 8 で求める．

$$H(e) = -\sum_{s\in S(e)} P(e,s)\log_2 P(e,s) \qquad (8)$$

ただし，手がかり表現を抽出する業績発表記事集合において，$P(e,s)$ は共通頻出表現 e が手がかり表現 s に係る確率，$S(e)$ は共通頻出表現 e が係る手がかり表現の集合である．閾値 T_e は，以下の式 9 によって設定する．

$$T_e = \alpha\log_2 Ns \qquad (9)$$

ただし，Ns は共通頻出表現を取得するのに使用した手がかり表現の数，α は定数 $(0<\alpha<1)$ である．

4.4 新たな手がかり表現の獲得

共通頻出表現の選別を行った後，その選別した共通頻出表現から新たな手がかり表現を獲得する．まず，抽出した共通頻出表現を含む文を抽出し，その中で共通頻出表現を含む節 P_a が係っている文節 P_b を獲得する．次に，節 P_a から共通頻出表現を除いた文字列を，文節 P_b に追加し，それを手がかり表現候補とする．また，新たな手がかり表現に対しても，様々な共通頻出表現が係っている手がかり表現は適切であるという仮定にもとづき，手がかり表現候補に対して共通頻出表現が係る確率に基づくエントロピーを式 10 で求め，閾値以上の候補を手がかり表現として抽出する．

$$H(s) = -\sum_{e\in E(s)} P(s,e)\log_2 P(s,e) \qquad (10)$$

ただし，手がかり表現を抽出する業績発表記事集合において，$P(s,e)$ は手がかり表現 s に対して共通頻出表現 e が係る確率，$E(s)$ は手がかり表現 s に係る共通頻出表現の集合である．閾値は，共通頻出表現と同様に式 9 によって設定するが，Ns は新たな手がかり表現を獲得するのに使用した共通頻出表現の数である．

4.5 不適切な手がかり表現，共通頻出表現の除去

上記の選別手法では，定数 α の値を下げ閾値を小さくしていくと不適切な手がかり表現や共通頻出表現が次第に多く取得される．そして，手がかり表現から共通頻出表現を獲得し，共通頻出表現から手がかり表現を獲得する処理を繰り返すため，不適切な手がかり表現が獲得されると，そこから新たな不適切な共通頻出表現が獲得されるという繰り返しになり，精度が急激に落ちていく．そこで，業績発表記事集合と同数の業績発表記事以外の文書集合（業績発表記事以外の文書集合は，2 章において業績発表記事と判定されなかった文書の集合とする.）を使用して，不適切な手がかり表現，および，共通頻出表現を除去する．これは，適切な手がかり表現は業績発表記事集合中に分散して出現し，不適切な手がかり表現は業績発表記事集合中にもそれ以外にも出現するという仮定にもとづく．具体的には，業績発表記事集合において手がかり表現のスコアを以下の式 11 で求める．

$$W(s,S_p) = P(s,S_p)H(s,S_p) \qquad (11)$$

ここで，$P(s,S_p)$ は業績発表記事集合 S_p における手がかり表現 s を含む文の出現確率，$H(s,S_p)$ は業績発表記事集合 S_p に含まれる各文書における手がかり表現 s の出現確率に基づくエントロピーであり，次式 12 で求める．

$$H(s,S_p) = -\sum_{d\in S_p} P(s,d)\log_2 P(s,d) \qquad (12)$$

ここで，$P(s,d)$ は文書 d における手がかり表現 s を含む文の出現確率を表す．また，業績発表記事以外の文書集合におけるスコア $W(s,S_n)$ も同様に求める．ただし，S_n を業績発表記事以外の文書集合とする．そして，手がかり表現 s の重み $W(s,S_p)$ が重み $W(s,S_n)$ の 2 倍より小さければ，その手がかり表現を除去する．同様の手法で共通頻出表現も除去する．なお，上記処理は，手がかり表現の抽出，共通頻出表現の抽出を行う度に行う．

4.6 手がかり表現からの業績要因の抽出

獲得した手がかり表現と共通頻出表現を使用して業績要因を抽出する．具体的には，4.2 節で述べた方法で手がかり表現に係る文節から派生する表現を取得し，その中の最長の表現の中に共通頻出表現が含まれている場合，その表現と手がかり表現を連結して業績要因とする．図 1 に，定数 α を 0.3 とした場合に獲得した手がかり表現と共通頻出表現を使用して抽出された業績要因をいくつか示す．なお，括弧の中は使用した手がかり表現である．

5 評価

本手法を実装し，評価を行った．実装にあたり，形態素解析器として ChaSen[2]，係り受け解析器として CaboCha[3]を

[2]http://chasen.aist-nara.ac.jp/hiki/ChaSen/
[3]http://chasen.org/~taku/software/cabocha/

- 液晶ディスプレー向けガラス基板の好調が寄与：「が寄与」
- 収益性が高いカードゲームのブーム一巡が響いた：「が響いた」
- 音楽ＣＤなどの販売が低迷した：「が低迷した」
- アニメＤＶＤ販売が伸びた：「が伸びた」

図 1: 抽出された業績要因 (括弧内は手がかり表現)

使用した．そして，01 年から 05 年の日経新聞記事集合から取得した 20880 個の業績発表記事から，手がかり表現，および，共通頻出表現を獲得し，業績要因を抽出した．まず，前処理である業績発表記事抽出を評価する．正解データは，テストデータから無作為に 10281 記事を抽出し，その中から人手で業績発表記事を抽出することで作成した．その正解データを使用して，業績発表記事抽出の精度，再現率を求めた結果，精度は 88.6%，再現率は 93.7% であった．

次に，業績要因抽出の評価を行った．正解データは，取得した 20880 個の業績発表記事の中から無作為に 243 個の記事を選び，さらに，その中から業績要因が記述してある文を人手で抽出して作成した．さらに，本手法によって獲得された手がかり表現と共通頻出表現を使用して 243 個の記事中から業績要因を獲得し，その業績要因が含まれている文を抽出した．そして，正解データと比較することで，精度，再現率を求めた．さらに，自動獲得した手がかり表現を人手で判定することで，手がかり表現の精度を求めた．表 1 に，本手法の定数 α を 0.6 から 0.2 まで変化させた場合の結果を示す．なお，F_s は手がかり表現獲得数，P_s はその精度，P_{cs}は 業績要因文の精度，R_{cs} は業績要因文の再現率である．比較手法として，交通事故事例に含まれる事故原因表現を抽出するために考案した手法 [3] を使用した．具体的には，本手法に対して，4.5 節で説明した不適切な表現の除去を行っていない手法である．ただし，比較手法の初期手がかり表現は「が好調」と「が不振」である．比較手法の結果を表 2 に示す．

表 1: 評価結果 (本手法)

α	F_s	$P_s(\%)$	$P_{cs}(\%)$	$R_{cs}(\%)$
0.6	8	100	100	2.0
0.4	139	92.1	93.8	15.3
0.3	922	78.7	79.1	62.8
0.2	3381	62.4	60.1	80.4

表 2: 評価結果 (比較手法)

α	F_s	$P_s(\%)$	$P_{cs}(\%)$	$R_{cs}(\%)$
0.6	10	100	100	6.4
0.4	59	67.8	96.3	6.6
0.35	938	65.8	78.4	62.5
0.3	1883	57.8	68.7	75.2
0.2	8078	26.8	51.1	91.9

6 考察

表 1 と表 2 を比較すると，閾値 0.3 の本手法の手がかり表現獲得数は 922 で精度が 78.7% であるのに対し，比較手法では閾値 0.35 で表現獲得数が 938 とほぼ同数でありながら精度が 65.8% であった．同一の獲得数では本手法の方が精度が高く，不適切な手がかり表現が抽出されていないことが分かる．比較手法では，手がかり表現から共通頻出表現を獲得し，共通頻出表現から手がかり表現を獲得する処理を繰り返すため，不適切な手がかり表現が獲得されると，そこから不適切な共通頻出表現が獲得されるという繰り返しになり，精度が大きく落ちていく．特に低い閾値ではその現象が顕著に表れ，閾値を低くすると不適切な手がかり表現が大幅に増えていった．それに対処するため，本手法では，不適切な手がかり表現を除去する手法を導入しており，表 1 より低い閾値でも高い精度を保っていることが分かる．

抽出した業績要因をみると，その業績要因が業績に対してポジティブかネガティブかを判定できる．例えば，図 1 の「液晶ディスプレー向けガラス基板の好調が寄与」はポジティブである．そして，本手法で獲得した業績要因がポジティブかネガティブかを自動的に判定できれば，本手法で獲得したデータは業績要因の分析だけでなく，全体的な景気動向の推定にも利用できる．すなわち，ある一定期間においてネガティブな業績要因数に比べてポジティブな業績要因数が多ければ，景気は上向くことが推定できる．そのため，今後の課題として，業績要因が業績に対してポジティブであるかネガティブであるかの判定を挙げる．具体的には，獲得した手がかり表現から推定できると考える．例えば「が好調」,「が寄与」という手がかり表現を含む業績要因はポジティブであろう．しかしながら,「売り上げが増加」はポジティブであるが,「有利子負債が増加」はネガティブであるので，共通頻出表現をも考慮して判定する必要があると考える．

7 まとめ

本研究では，経済新聞記事から企業の業績発表に関する記事を抽出し，その要因を自動的に抽出する手法を提案した．抽出方法は，業績要因が記述してある部分を示す手がかり表現を自動的に獲得することで行うが，少数の手がかり表現を入力することで自動的に多数の手がかり表現を取得する手法であり，他タスクへの適用も可能である．評価の結果，交通事故事例に含まれる事故原因表現を抽出するために考案した手法 [3] に比べ低い閾値においても手がかり表現の精度の低下が少なく，良好な結果を得た．今後の課題として，業績要因が業績に対してポジティブであるかネガティブであるかの判定を挙げる．

謝辞

本研究の一部は，文部科学省科学研究費特定領域研究 (B)(2)16092213，及び，２１世紀 COE プログラム「インテリジェントヒューマンセンシング」(豊橋技術科学大学) の援助により行われた．また，言語データとして，日経新聞 CD-ROM の使用を許可して頂いた日本経済新聞社に深謝する．

参考文献

[1] 池原悟, 宮崎正弘, 白井諭, 横尾昭男, 中岩浩巳, 小倉健太郎, 大山芳史, 林良彦 (編): 日本語語彙大系, 岩波書店 (1997).

[2] 那須川哲哉, 金山博, 坪井祐太, 渡辺日出雄: 好不評文脈を応用した自然言語処理, 言語処理学会第 11 回年次大会発表論文集, pp. 153–156 (2005).

[3] 酒井浩之, 梅村祥之, 増山繁: 交通事故事例に含まれる事故原因表現の新聞記事からの抽出, 自然言語処理, Vol. 13, No. 4, pp. 99–124 (2006).

[4] 峠泰成, 山本和英: 手がかり語自動取得による Web 掲示板からの評価文抽出, 言語処理学会第 10 回年次大会発表論文集, pp. 107–110 (2004).

[5] Vapnik, V.: *Statistical Learning Theory*, Wiley (1999).