1F4-01

キー概念辞書を用いないテキストの自動分類

An Automatic Classification of Textual Data without using a Key Concept Dictionary

櫻井 茂明
Shigeaki Sakurai

酢山 明弘
Akihiro Suyama

(株)東芝 研究開発センター
Corporate Research & Development Center, Toshiba Corporation

We proposed a method that acquires automatically concept relations described by the format of a fuzzy decision tree. The method uses a key concept dictionary created by a human expert. The human expert has to create the key concept dictionary through trial and error. We need a learning method that does not need a key concept dictionary. The paper investigates the effects of adjusted learning parameters and pruned fuzzy decision trees for the learning method. The paper also compares a method based on key concept dictionaries with a method based on words through e-mail classification tasks.

1. はじめに

テキストデータが簡便に収集できるようになり、テキストマイニング研究が活発化している[1][4][5]。我々の研究グループにおいても、テキストに含まれる表現を階層的にまとめたキー概念辞書と、キー概念の組み合わせが示す意味を記述する構造抽出ルールに基づいて、テキストを分類し、その傾向を視覚的に表示するテキストマイニングシステムを提案[2]する一方、キー概念辞書の構築を支援するツール(CADDIE)[3]や構造抽出ルールを帰納学習する方法[6]を研究開発している。CADDIE を導入することにより、キー概念辞書を効率的に構築できるようになるものの、タスクごとに、キー概念辞書を生成する必要は依然としてあり、多くのタスク向けにテキストマイニングシステムを適用することは困難であった。

このため、現在、キー概念辞書を利用しないテキストマイニングシステムの構築に向けた研究開発を行っている。本論文では、その手始めとして、キー概念辞書を利用する代わりに、形態素解析した単語を利用することにより、構造抽出ルールを学習する方法を取り上げ、学習パラメータを調整した場合の効果及び MDL 基準に基づいた枝刈り法を導入した場合の効果を検証する。これにより、キー概念辞書を利用しない構造抽出ルールの学習法の可能性を検討する。

2. テキストマイニング

2.1 テキストマイニングの流れ

論文[2]に提案したテキストマイニングシステムは、「形態素解析」、「情報抽出」、「構造抽出」、といった各処理を順次実施することにより、テキストを分類する。すなわち、キー概念辞書を参照することにより、形態素解析された単語集合の中から、「情報抽出」がキー概念を抽出し、構造抽出ルールを参照することにより、「構造抽出」がキー概念の組み合わせが示す意味(テキストクラス)を抽出する。ここで、キー概念辞書は、概念クラス、キー概念、表層表現といった 3 階層で記述されており、正規表現を用いて、実際にテキストに含まれる表現が表層表現に記述されている。また、構造抽出ルールは、キー概念の組み合わせを条件部、テキストクラスを結論部に持つように記述されている。本テキストマイニングシステムは、各テキストに割り当てられるテキストクラスに基づいてテキストを分類し、分類されたテキストに関する統計情報を利用者に提示する。これにより、利用者は、興味あるテキストだけを読むことができ、全体的な傾向を把握することができる。

連絡先:櫻井 茂明, 〒212-8582 神奈川県川崎市幸区小向東芝町1 (株)東芝 研究開発センター 知識メディアラボラトリー, Tel: 044-549-2398, shigeaki.sakurai@toshiba.co.jp

2.2 構造抽出ルールの学習・推論

構造抽出ルールの学習法として、キー概念を利用する方法、単語を利用する方法、単語のグループを利用する方法、連続する単語の組み合わせ(キーフレーズ)を利用する方法、を提案している。各手法においては、形態素解析を利用して、テキストを単語に分解し、分解された単語に基づいて属性ベクトルを生成する。このとき、属性及び属性値としてどのような値を取るかによって、各手法に違いが存在する。キー概念辞書を利用する方法では、概念クラスを属性、キー概念の集合を属性値とみなし、単語を利用する方法では、単語を属性、単語が出現したかどうかを属性値とみなしている。この属性ベクトルと、利用者がテキストに付与したテキストクラスを組にすることにより、訓練事例を生成する。このようにして生成した訓練事例の集合を、ファジィ帰納学習アルゴリズム IDF に適用し、ファジィ決定木形式の構造抽出ルールを学習する。

ファジィ決定木を用いた推論では、訓練事例の場合と同様にして、テキストに対応する属性ベクトルを生成し、属性ベクトルに基づいて、ファジィ決定木の最上位のノード(ルートノード)から、順次下位のノードへと推論を行っていく。すなわち、分岐ノードにおいては、分岐ノードに割り当てられている属性で、属性ベクトルの属性値を評価することにより、確信度を更新した属性ベクトルを下位のノードに伝播する。末端ノードにおいては、伝播した属性ベクトルの確信度と末端ノードに割り当てられている分類クラスの確信度で、分類クラスの評価を行う。最終的には、各末端ノードにおける評価結果をまとめることにより、確信度付の分類クラスを推論する。通常、最大の確信度を与えるひとつの分類クラスをテキストに対応するテキストクラスとして推論する。

3. 単語ベースの学習法の改良

3.1 学習パラメータの調整

構造抽出ルールの学習に利用する IDF には、枝刈り率、分割率、といったパラメータが存在し、各パラメータは、式(1)、式(2)を用いて定義される。

$$\text{枝刈り率} = \frac{\text{対象ノードにおける最大の確信度を与える}\\ \text{クラスを持つ事例の確信度の和}}{\text{対象ノードに伝播する事例の確信度の和}} \quad (1)$$

$$\text{分割率} = \frac{\text{対象ノードに伝播する事例の確信度の和}}{\text{学習に利用する事例の確信度の和}} \quad (2)$$

従来は、このパラメータの値として、IDF のデフォルトのパラメータである、枝刈り率=1.00 及び分割率=0.00 を利用した学習を行っていた。デフォルトの値の場合、事例を最大限に細かく分解することにより、訓練事例に対する正解率が最も高くなる構造抽出ルールを学習する。しかしながら、このような構造抽出ルールは、過学習される傾向にあり、必ずしも新たなテキストを分類する上での、適切な構造抽出ルールにはなっていない。このため、学習パラメータを適切に調整することにより、分類精度の向上を期待できる。そこで、本論文では、学習パラメータを変化させて学習した構造抽出ルールに対して、学習に利用しなかった事例を評価し、最大の正解率を与える構造抽出ルールの生成に利用されたパラメータを、当該学習事例に対するパラメータとする調整を行う。

3.2 枝刈り法の導入

過学習の問題に対しては、一旦学習した決定木から不要な枝を刈り込んだ決定木を利用することにより、新たなデータに対する分類精度が向上することが期待されている。ファジィ決定木においても、同様の効果を期待することができるので、ファジィ決定木における枝刈り法を検討する。枝刈り法としては、一様分布に基づく方法、MDL を利用する方法、等が提案されているが、本論文では、MDL に基づいた方法を検討する。ただし、構造抽出ルールにおいては、離散値あるいはその集合に、属性値が限定されるので、離散的なものに限定して検討を行う。また、以下においては、S_B は分岐ノードの集合、S_T は末端ノードの集合、S_A は属性の集合、S_C は離散分類クラスの集合、V_{A_b} は離散分岐ノード b に対応する属性 A_b のすべての基本属性値(ひとつのキー概念だけからなる属性値)の集合、v_{A_b} は離散分岐ノード b に対応する属性 A_b の枝に現れる基本属性値の集合、n_t は末端ノード t に伝播する事例の数、x_t は末端ノード t における最大確信度を与えるクラスと事例のクラスが一致しない事例の数、$|\bullet|$ は集合に含まれる要素数を求める演算とする。

訓練事例の集合を表現するファジィ決定木を記述する構成要素としては、木構造の表現、分岐ノードの表現、末端ノードの表現、例外事例の表現、といった要素が考えられる。そこで、以下においては、各要素について順次検討していく。

木構造の表現は、ノードの巡回方法を決定することにより、ノードの位置関係を保持したまま、一列に並べることが可能である。また、各ノードは、分岐ノードか、末端ノードのいずれかであり、ひとつのノードを表現するのに、1 ビット必要である。このため、木構造を表現するには、ファジィ決定木を構成するノード数($|S_B|+|S_T|$)と等しいビット数が必要になる。

分岐ノードの表現には、分岐ノードに割り当てられる属性、分岐ノードにつながる下位のノードへの枝に割り当てられる基本属性値といった要素が含まれている。分岐ノードに割り当てられる属性は、選択可能なすべての属性の中から選ぶことができるため、どの属性が割り当てられるかを表現するには、属性数をビット表現するだけのビット数($\log_2|S_A|$)が必要になる。一方、IDF においては、学習時に現れた基本属性値に対応する枝だけを生成するため、どの基本属性値が出現したかを表現する必要がある。枝の数($|v_{A_b}|$)が決定された場合、この表現の組み合わせは、${}_{|V_{Ab}|}C_{|v_{Ab}|}$ 通り存在し、その中のひとつの組み合わせが選ばれることになる。このため、どの基本属性値が出現したかを表現するには、この組み合わせをビット表現するだけのビット数($\log_2 {}_{|V_{Ab}|}C_{|v_{Ab}|}$)及び枝の数をビット表現するだけのビット数($\log_2|v_{A_b}|$)が必要となる。このため、分岐ノードを表現するには、これらのビット数を加えた値が必要となる。

末端ノードを表現するには、末端ノードに割り当てられている確信度付きの分類クラスを表現する必要がある。確信度は(0,1]の実数値として与えられており、この値を正確にビット表現することはできない。ある程度の桁数を決めて、確信度を表現することは可能であるものの、木構造や分岐ノードに比べるとその記述長はかなり大きくなると予想され、記述長が末端ノードの表現に過度に依存すると考えられる。そこで、通常の決定木における記述長の場合と同様に、ひとつの分類クラスで末端ノードを代表させ、その記述長を末端ノードの記述長とする。従って、末端ノードを表現するには、分類クラスの数を表現するだけのビット数($\log_2|S_C|$)が必要となる。

IDF に基づいたファジィ決定木の学習では、確信度の付いた訓練事例が確信度の付きの分類クラスを持つ末端ノードに伝播する。このため、例外事例を正確に表現するには、訓練事例の確信度を表現するとともに、訓練事例の持つ分類クラスと、末端ノードが持つそれ以外の分類クラスを、確信度を含めて表現する必要がある。このような表現は、末端ノードの表現の際に検討したように、例外事例の表現に過度に依存した記述長を導くことになる。そこで、末端ノードにおける最大確信度を待つ分類クラスと一致しない分類クラスを例外事例とみなすとともに、訓練事例の確信度は表現しないことにする。このため、末端ノードを表現するには、伝播した訓練事例のうち、どの訓練事例が例外となり、どういった分類クラスを持つかを表現するだけでよい。すなわち、n_t 個の事例が伝播し、例外事例が x_t 個ある場合には、この組み合わせは、${}_{|n_t|}C_{|x_t|}$ 通り存在するので、この組み合わせをビット表現するだけのビット数($\log_2 {}_{|n_t|}C_{|x_t|}$)が必要になる。この他、各例外事例の数を表現するビット数($\log_2(x_t+1)$)及び、各例外事例がどのクラスを持つかを表現するビット数($\log_2(|S_C|-1)$)が必要になる。

以上により、各構成要素を記述する記述長は、表 1 のように与えられる。

表 1:ファジィ決定木の記述長

木	$\|S_B\|+\|S_T\|$
分岐	$\|S_B\|\log_2\|S_A\|+\sum_{b\in S_B}(\log_2 {}_{|V_{Ab}|}C_{|v_{Ab}|}+\log_2|v_{A_b}|)$
末端	$\|S_T\|\log_2\|S_C\|$
事例	$\sum_{t\in S_T}\{\log_2 {}_{n_t}C_{x_t}+\log_2(x_t+1)+x_t\log_2(\|S_C\|-1)\}$

このとき、最小の記述長を与えるファジィ決定木を正確に選択するには、生成可能なすべてのファジィ決定木に対応する MDL を計算し、その値に基づいて、最小の記述長を与えるファジィ決定木を選択する必要がある。しかしながら、計算量を考えた場合、このようなすべてのファジィ決定木を生成する方法は現実的な方法ではない。そこで、デフォルトのパラメータで一旦ファジィ決定木を学習し、記述長が改善する限り、冗長な枝を刈り込んで行くことにより、MDL を基準としたファジィ決定木を生成

する。このようなファジィ決定木は、訓練事例に対する過度の依存を避けることができるので、新たなテキストに対して、高い分類精度を与えることが期待できる。

4. 数値実験

4.1 メール分類システム

近年、コールセンターには、多数の電子メールが顧客から送られており、その数は増大する傾向にある。このような電子メールは、顧客満足度の向上を図る上での重要なデータになると考えられるものの、必ずしも十分な分析が行われていない。それればかりか、その数の増大に伴って、分析はより困難なものになっている。

この問題に対して、電子メールを構成するデータが主にテキストデータから構成されることに着目し、テキストマイニング技術を利用した電子メールの分析が検討されている。このような分析が実現できれば、顧客満足度の向上へと結び付く分析を行うことが期待できる。本論文では、電子メールデータを構成するデータのうち、タイトル、テキストとして記載された本文を、テキストマイニングシステムへの入力とし、利用者が与えたテキストクラスに基づいて、電子メールの自動分類を行う。これにより、利用者は、興味のあるクラスに含まれる電子メールを詳細に分析し、各クラスに含まれるデータ件数の傾向を把握することができる。

4.2 実験データ

実験においては、東芝のコールセンターが収集した電子メールデータのうち、466 件のデータを利用した実験を行う。本データに対しては、製品分類、内容分類といった 2 種類の分類がオペレータによってなされており、各クラスには、表 2 に示す件数の電子メールが存在する。これらのデータには、個人情報が含まれており、その取り扱いに注意しなければならない。そこで、個人情報に該当する部分を特定の文字列に変換したデータを実験では利用する。

表 2:テキストの分布

(a)製品分類

クラス	データ数
洗濯機	103
掃除機	81
冷蔵庫	84
レンジ	153
その他	45

(b)内容分類

クラス	データ数
質問	266
要求	93
提案	10
苦情	83
その他	14

本データに対して。形態素解析を実施し、予備実験の結果得られたしきい値(0.005)以上の tfidf 値を持つ単語を抽出する。このとき、2,098 個の単語が実験データから抽出されるため、単語ベースの構造抽出ルールの学習においては、2,098 個の属性が存在する。また、各属性値は、単語が存在するかどうかの 2 値で与えられる。一方、本データに記載されている内容を分析することにより、キー概念辞書が生成されており、キー概念辞書のサイズ及びその生成時間が、表 3 に与えられている。キー概念辞書に基づいた構造抽出ルールの学習においては、概念クラスが属性、キー概念の集合が属性値とみなされる。このため、製品分類においては、20 個の属性、2^{235} 個の属性値が与えられ、内容分類においては、35 個の属性、2^{96} 個の属性値が与えられる。

表 3:キー概念辞書

	製品分類	内容分類
概念クラス	20	35
キー概念	235	96
表層表現	427	259
生成時間	17H	24H

4.3 実験方法

実験においては、10 cross validation に基づいた実験を行う。すなわち、466 件のテキストの集合を 10 個の部分集合に分割し、このうちの 9 つの部分集合を学習に利用するデータとし、残りのひとつを評価用のデータとする。この学習用のデータを、構造抽出ルールの学習法に適用することにより、ファジィ決定木形式の構造抽出ルールを学習し、学習した構造抽出ルールを用いて、評価用のデータに対応するテキストクラスを評価する。このとき、評価用のデータに割り当てられているテキストクラスと、推論されたテキストが一致するかどうかを評価することにより、式(3)で定義される正解率を計算し、学習した構造抽出ルールの分類精度を評価する。このような学習と評価を評価用のデータを順次変えることにより 10 回実施し、各実験における正解率及び 10 回の実験における平均正解率を評価する。

$$\text{正解率} = \frac{\text{分類クラスが一致する評価データの数}}{\text{評価データの数}} \qquad (3)$$

以上に示した 10 cross validation に基づいた実験を、キー概念辞書に基づいた学習方法及び単語ベースに基づいた学習方法の各手法に対して、パラメータ調整を行った場合、枝刈り法を導入した場合に対して行い、各手法の効果を検証する。

4.4 実験結果

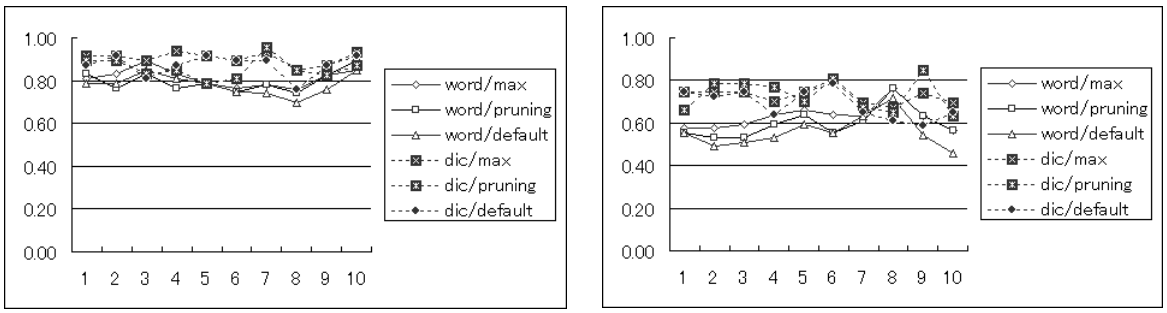
図 1 に各手法及び各分類基準に対する正解率が変化する様子を示す。図においては、x 軸が 10 cross validation における 1 回の実験を表しており、y 軸が正解率を表している。また、表 4 は、10 回の実験における正解率の平均値を表している。ここで、図 1、表 4 においては、word が単語ベースの結果、dic がキー概念辞書を利用した場合の結果、max がパラメータ調整を行った場合の結果、pruning が枝刈り法を導入した場合の結果、default が IDF のデフォルトのパラメータを利用した場合の結果を表している。

4.5 考察

キー概念辞書の効果:

実験結果から分るように、製品分類、内容分類のいずれの場合においても、キー概念辞書を利用した学習法の分類精度が高くなっている。キー概念辞書においては、表現の異なる単語であったとしても、同じ意味を示す表現を同一のキー概念としてまとめ上げており、単語ベースのものに比べて、テキストに適切な特徴付けを行うことができる。また、形態素解析の誤りによる不適切な分割が行われた場合にも、正しい分割に基づいた単語の組み合わせで特徴付けを行うことができ、複合語に対応した特徴付けも行うことができる。このため、キー概念辞書を利用した場合の分類精度が高くなったと考えられる。

また、キー概念辞書を利用した場合には、対象とする属性の数が、単語ベースのものに比べて格段に少なくなっている。このため、構造抽出ルールを高速に学習することができる。一方、構造抽出ルールを用いた推論では、推論時間の大部分は形態



(a) 製品分類　　(b) 内容分類

図 1: 正解率の推移

表 4: 平均正解率

(a) 製品分類

	max	pruning	default
word	81.3%	79.6%	78.1%
dic	90.6%	85.6%	87.1%

(b) 内容分類

	max	pruning	default
word	62.7%	59.7%	55.8%
dic	73.0%	73.2%	68.9%

素解析に必要な時間となるため、どちらの方法であったとしても、大きな違いは観測されない。

パラメータ調整の効果:
パラメータ調整を行うことにより、デフォルトパラメータを利用するよりも、平均正解率が 3% ～ 7% 程度向上している。特に、内容分類の場合に、その効果が高くなっている。当初予想した通り、デフォルトパラメータでは、過学習が行われており、パラメータ調整により、過学習された枝の生成が回避できたため、平均正解率が向上したと考えられる。また、内容分類の場合には、複数の単語を組合わせた表現が、分類を行う上での重要な表現となる傾向にあり、単語ベースの場合に、事例の少なさから、このような組合わせが正しく選択されないことがある。このため、デフォルトのパラメータで学習した場合に、過学習される枝が多くなり、パラメータ調整の効果が大きくなったと考えられる。

枝刈り導入の効果:
枝刈りを行った場合の平均正解率は、多くの場合において、デフォルトパラメータを利用する場合よりも高くなっており、MDL 基準が、過学習した枝の判別に、ある程度効果があることが確認できた。しかしながら、最大の正解率を与える場合に比べれば、分類精度が劣っており、ファジィ決定木の記述方法、枝刈り対象ノードの決定方法等に、改良する余地があると考えられる。

以上に議論したように、パラメータ調整、枝刈り法の導入は、各手法における分類精度の向上に効果がある。しかしながら、これらの効果は、単語ベースの方式に限ったものにはなっておらず、キー概念辞書を利用しない学習法を確立する上では、まだ十分な成果が得られているとはいえない。単語ベースの方式に関連した改良が一層必要と考えられる。

5. まとめと今後の課題

本論文では、構造抽出ルールの学習法の改良として、パラメータの調整、枝刈り法の導入を検討し、その効果を、コールセンターメールの分類問題に適用して検証した。また、単語ベースの学習法における分類精度を明らかにし、キー概念辞書を利用しない学習法の改良に向けてのベースとなる分類精度を得ることができた。

今後の課題としては、キー概念辞書を利用しない学習法の確立に向けて、頻出する単語のつながり、単語の共起情報、品詞情報、等の利用による分類精度の向上が期待できるので、それらの利用方法を検討し、効果を検証していきたい。

参考文献

[1] R. Feldman, I. Dagan, and H. Hirsh: “Mining Text using Keyword Distributions,” Journal of Intelligent Information Systems, **10**, 3, 281-300 (1998).

[2] 市村 由美, 中山 康子, 赤羽 俊男, 三好 みよ子, 関口 寿一, 藤原 庸祐: 「営業日報を対象としたテキストマイニング-成功事例及び機会損失情報の抽出-」, 人工知能学会全国大会 (第 14 回) 論文集, 532-534 (2000).

[3] 市村 由美, 酢山 明弘, 櫻井 茂明, 折原 良平: 「知識辞書構築支援ツールの開発」, 情報処理学会研究報告, 2001-NL-143, 25-31 (2001).

[4] 町田 明子, 林 俊克: 「ネット上書き込み情報のテキストマイニング」, 第 4 回日本感性工学会大会予稿集 2002, 249 (2002).

[5] 舘野 昌一: 「テキスト感性表現の抽出によるお客様の声の活用方法」, 第 4 回日本感性工学会大会予稿集 2002, 242 (2002).

[6] 櫻井 茂明, 酢山 明弘: 「ファジィ帰納学習におけるキー概念集合を含む属性値の扱い」, 日本ファジィ学会誌, **14**, 6, 640-647 (2002).