リスク抽出タスクにおけるルールとSVMの相補利用

平田勝大[1)]　梅村恭司[1)]　関根聡[2)]
豊橋技術科学大学 情報工学系[1)]
ニューヨーク大学[2)]

概要

リスク抽出タスクとは、有価証券報告書からリスク情報を抽出するタスクである。上場企業等が年度ごとに公表する有価証券報告書は投資家にとって非常に重要な書類であるが、数多くの企業から膨大な文書として提出されるため、そこから必要な情報を読み取ることは多大な労力が必要になる。そこで、投資家が特に興味を持つ情報について、構造化された形で情報が提示されると便利である。そのような情報の一つに企業が抱えるリスクの情報があり、今回の研究では、リスクの情報を文書から抽出し、その種類によって分割するタスクを対象にする。すでに有価証券報告書の中で箇条書きのような形で半構造化されていない文章を対象にSVMを利用してシステムを構築した。

1 はじめに

有価証券報告書は、上場企業等が年度ごとに事業の状況、財務状態や経営成績等を記載して公表されている書類である。この書類に記載されている情報は、投資家が投資をする際には重要な判断材料となっている。また、これらの情報は、投資家のみならず、一般的にも有益な情報である。

有価証券報告書には、100 ページ以上にもわたって企業情報が詳細に記載されており、上場企業のみでも 4500 社以上あるため、これら書類をすべて調べることは手間がかかる。このため、これらの書類から文書すべてを見ることなく情報を得ることができれば、企業について調べる上で有用である。

文書すべてを見ることなく文書の内容を見るには、必要な文書のみを抽出することが有効ある。今回は、リスク抽出タスクとして「事業等のリスク」の項目からリスク情報を抽出する。具体的には、リスクについて述べている文を抽出し、リスク内容ごとに分割するタスクである。このタスクは、文書分類のタスクであり、実際に需要のある処理であるが、分類するための文書の区切りが不明であり、標準的な分類手法[4,5,6]で分類できるかもわからない。

「事業等のリスク」の文書には、箇条書き表現のある文書と箇条書き表現のない文書がある。箇条書き表現のある文書では、ルールによってリスク内容ごとに分割した文書を比較的抽出しやすいが、箇条書き表現のない文章では難しい。本研究では、この箇条書き表現のない文章からの抽出方法を検討したことを報告する。分類は標準の方法を使用し、機械学習法であるサポートベクターマシン(SVM)を利用する。

初めに、句点で区切られる文書をひとつの文、各文を形態素解析した単語の出現頻度を特徴ベクトルとして、各文がリスクについて述べている文であるかを分類する。次に分類した文を順番に結合し、結合部分における前後の文の単語出現頻度を特徴ベクトルとして、どの 2 文で内容が分割されるかを分類する。以上の処理を行い、有価

証券報告書からリスク情報を内容ごとに分割して抽出したことを報告する。

2 有価証券報告書

有価証券報告書とは、有価証券である株券や債券を使って 1 億円以上の資金調達をする企業や株式を証券取引所などに上場や公開している企業が年度ごとに提出を義務付けられている書類である。この書類には、年度ごとの企業の事業内容や一年間の業績、設備投資の状況等が記載されており、本研究では、特に「事業等のリスク」の項目からリスク情報を抽出する。

3 サポートベクターマシン

サポートベクターマシンとは、2 つのクラスを識別する識別器を構成するための学習法である。学習データを用いて、2 つのクラスを線形分離し、分離超平面と最も近い学習データとの距離が最大になるようなモデルを学習する。この手法により、線形分離できるデータにおいては、未学習のデータにおいても高い識別性能を持つ学習手法の一つである。本研究では、SVM ツールである SVMlight[2]を利用する。

4 形態素解析

日本語は文書中の単語が区切られていないため、どの部分文字列が単語であるか判定する必要がある。形態素解析は、文字列の単語候補を辞書から調べ、単語の品詞の接続情報を用いて単語候補から単語を判別する解析手法である。この手法を用いることで、意味を持つ最小単位である単語に分割することができる。本研究では、形態素解析ツールである茶筌[3]を使用して、文書を単語に分割する。

5 リスク情報の抽出

箇条書き表現のない文書に対するリスク情報の抽出には、2 つの手順を用いて抽出する。それぞれの手順は、データを調査して決定した。手順 1 で、文書を句点によって分割した文がリスクに関する文であるかを SVM でテキスト分類する。手順 2 では、手順 1 で分類した文のうち、ある 2 つの文の間で内容が分割されるとして、内容が分割される文がどの文であるかを SVM で分類する。

○手順 1 リスクに関する文書の分類

初めに有価証券報告書の“事業等のリスク”の文書を句点で分割し、分割された文書がリスクに関する文書であるか SVM を利用して分類する。SVM の学習データには、500 企業に対して手作業で抽出したデータを使用する。SVM の特徴ベクトルには、学習データを形態素解析で分割した単語のうち、ひらがなを含む単語は 4 文字以上の単語、含まない単語は 2 文字以上の単語を合わせたものから出現頻度順上位 3000 語を使用する。

○手順 2 内容が分割される文の分類

手順 1 の文書を結合し、内容が分割される位置の前後の文書を SVM によって分類する。SVM の学習データには、手順 1 で使用したデータと同じものを使用し、特徴ベクトルには、手順 2 と同様の単語を利用し、前後 2 つの文書で別にした出現頻度を利用する。

6 抽出結果

上場企業の有価証券報告書のうち、箇条書き表現を利用したルールでは処理できない 405 企業の文書から、SVM を利用してリ

スク情報を抽出した。抽出結果と学習データとは異なるデータを実際に見て作成した50企業の正解データと比較した。この結果を表1に示す。

	再現率	適合率
リスクに関する文の分類	97.3	91.5
内容が分割される文の分類	78.6	65.7
全体の抽出結果	60.9	64.8

表1.リスク情報の抽出

表1より、内容が分割される文の分類の性能に問題があり、全体の抽出結果に大きな影響を与えていることがわかる。

次に、リスクに関する文の分類、内容の分割位置の判定における特徴的な単語の例を表2,3に示し、抽出例を図1,2に示す。特徴的な単語とは、学習データにおいて、正例に多く出現し、負例にあまり出現しない正例を特徴づける単語と、負例に多く出現し、正例にあまり出現しない負例を特徴づける単語である。

図1,2の例のうち、[ ]内が抽出結果である。図1の下線部ようにリスクに関する文であるかないかの分類は、分類しやすい文が多いが、リスクの概要を述べる文や図2の下線部のような「これら」などの指示語を含む文はうまく分類できない場合があった。また、内容が分割される文の分類は、

正例	負例
賠償	本書
取扱	文中
前提	網羅
改正	経理
地震	事項

表2.リスクに関する文の特徴単語

分割位置の前の文		分割位置の後の文	
正例	負例	正例	負例
リスクヘッジ	小規模	算出	リスクヘッジ
報告	公的	電機	報告
部長	化粧	電波	部長
人件	台湾	役割	人件
したがって	独占	運転	したがって

表3.内容の分割位置の特徴単語

本項目において将来に関する事項が含まれておりますが、当該事項は当連結会計年度末現在において判断したものであります。[当グループは橋梁の設計、製作、架設工事などを主な業務としておりますが、発注元は国土交通省や日本道路公団などの公団・公社が多く、公共関連事業に大きく依存しております。したがって、公共関連事業が伸び悩む昨今の情勢下では、当グループの売上高が減少する可能性があります。]

図1.抽出例1

[当社の仮設資材を使用する建設現場は、土木・建築工事ともに大型工事現場の場合が多く、売上に占める大型工事現場の割合は大きなものとなっております。さらに、首都圏地区を除く地方の大型工事は公共事業が大半を占めているため、引続く公共事業予算縮減の中、予想を上回る公共事業の削減が行われた場合には、少なからず業績に影響を及ぼす可能性があります。][なお大型工事現場向けの売上計上時期は各四半期必ずしも平準化したものとはならず、事業年度後半に集中する傾向があります。また、取引先の民事再生、破産等により債務不履行が発生した場合には、業績に影響を及ぼします。] なお、これらの他にも、購入資材価格の上昇、株価水準等、様々なリスクが存在しており、ここに記載したリスクが当社の全てのリスクではありません。

図2.抽出例2

図 1 のように直前の分をまとめる「したがって」のような単語が特徴的であり、分割位置の前後の文の両方で特徴的な単語となっている。この特徴は、表 3 の特徴単語において分割される文の前の文の正例と後の文の負例が同様になることからもわかる。

7 エラー分析

抽出例のように、直前までの文をまとめる表現の文は、その文の直前で内容が分割されず、その文の直後で内容が分割されやすいことを分類しやすい。しかしながら、直前までの文をまとめる表現でない文の場合は、特徴的な単語があまりないため、内容が分割される文の分類の性能が下がってしまったと考えられる。これは、前後の文中の単語の関係を考慮する必要があり、この関係は単純に等しい単語が出現するわけではないので難しい。また、企業ごとに内容の分割基準が異なっている場合があり、この場合は内容の分割位置の前後の文だけではわからない。そのため、はなれた位置にある文脈の流れを考慮する必要があるが、これは実装が困難である。

8 まとめ

箇条書き表現を利用したルールではリスク抽出できない文書に対して、SVM でリスク抽出をすることを行った。どの文がリスク情報に関する文であるかの分類は、特徴的な文が多く、SVM でうまく分類できる場合が多いが、内容が分割される文の分類は難しいことがわかった。この方法でルールと SVM で相補利用したところ、ルールで処理しにくい文書が処理されたことを報告する。

謝辞

本研は文部科学省 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」の援助により行われた。

参考文献

[1]栗田多喜夫:サポートベクターマシン入門, 産業技術総合研究所 脳神経情報研究部門 (2002)

[2]Thorsten Joachims : Support Vector Machine, Cornell University (2004)

[3]松本裕治:日本語形態素解析システム『茶筌』 version 2.2.1 使用説明書 (2000)

[4]Christopher D. Manning and Hinrich Schuetze. : Foundations of Statistical Natural Language Processing,The MIT Press (1999)

[5]Ian H. Witten and Eibe Frank : Data Mining,Morgan Kaufmann Publishers (2000)

[6]北研二,津田和彦,獅々堀正幹 : 情報検索アルゴリズム,共立出版(2001)