Support Vector Machine (SVM) を用いた自然文読影レポートからの医学的知識の抽出

安永　晋　川上 洋一　笹井　浩介

安藤 祐斗

富山県立大学　電子情報工学科

October 30, 2020

はじめに

背景

近年、病院内の情報システム化が進み、大量のデータを蓄積することが可能になっている。過去の自然読影レポートから効率的に医学的知識を抽出し、これを用いて入力支援情報を提示するレポーティングシステムが開発されている。

目的

レポーティングシステムで扱う過去の自然読影レポートは、基本的に自然文なのでそのままでは医学的知識を抽出するのは困難である。よって、自然文に形態素解析、構造化処理を行い、医学的知識を抽出し易くする必要がある。本論文では、SVM（Support Vector Machine）と呼ばれるアルゴリズムを適用し、自然文から医学的知識を抽出する技術と実験結果について述べる。

SVM(Support Vector Machine) とは

SVM について

SVM とは、過去の事例の学習によってベクトルを二値あるいは多値に分類する方法である。m 個の n 次元事例ベクトル $X_1,..,X_m$ を正負いずれかのクラスに属しているとすると、n 次元ベクトル W と定数 b を適切に選ぶことによって分離超平面 (実線) と超平面 (点線) は次式を満たす。

$$(w \cdot x) + b = 0$$
$$(w \cdot x) + b \pm 1$$

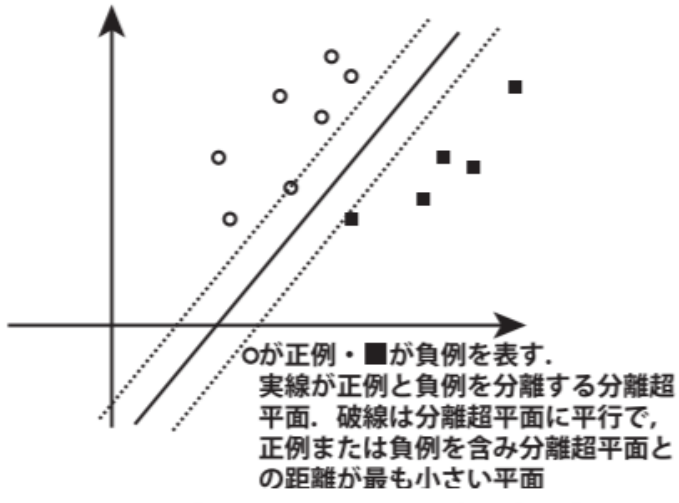

図 1: SVM

SVM(Support Vector Machine) とは

SVM について

・2 つの超平面の距離は $\frac{2}{\|w\|}$ で表され、これを最大にすればよい。
・$(w \cdot x_1) + b \geq 1$ と $(w \cdot x_1) + b \leq 1$ の制約条件のもと w と b を求める。
・w および b を求める処理が事例による学習である。

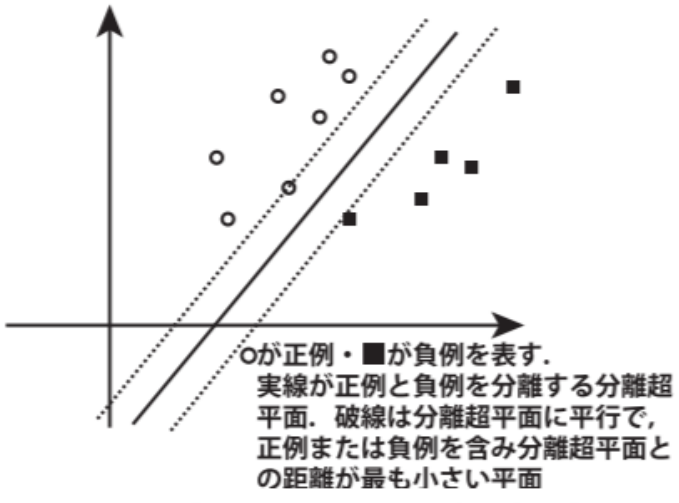

図 2: SVM

SVM の医学的知識抽出への応用

SVM を構造化処理に用いる具体的な方法

単語（見出し語・品詞の種類・語の長さなど）の情報をベクトルに変換し、このベクトルが特定の属性を持つかどうかを求める際に用いられる。

SVM を用いた言語処理ツール「YamCha」

文に含まれる単語および前後いくつかの単語の情報（品詞やひらがな・カタカナ・漢字などの文字種、単語の長さ、属性）をベクトルに変換し、それに対して SVM を用い、特定の属性を持つ単語・フレーズを抽出する。
（例）新聞記事を用いて学習を行い、これをもとに文章中から「地名」「人名」「時間」「日付」などを抽出する。

実験

実験概要

・頭部 MR に関する読影レポートから、医学的知識を抽出する。
・読影レポートにて、文は大きく分けて特徴記述文、診断記述文の 2 種類に分けられ、抽出すべき医学的知識は、「撮影条件」「部位」「特徴」「基本所見」「結語 (特徴記述文)」「診断」「結語 (診断記述文)」とする

特徴記述文と診断記述文

・特徴記述文例:FLAIR 像にて基底核に高信号域を認める。
「撮影条件」にて「部位」に「特徴」の「基本所見」を「結語」

・診断記述文例:急性期のラクナ梗塞を疑う
「診断」を「結語」

実験

実験ツール

・CaboCha(自然文から形態素解析・タグ付け・文節区切り・係り受け解析を行う)

実験対象

学習用データ：　レポート 270 件、文の数 408 文 (兵庫医科大学院)
処理対象データ：　レポート 74 研、文の数 262 文（大阪大学付属病院）

実験

実験手順

Step1　学習用データに形態素解析を行い、タグをつける。
Step2　タグ付けされたデータをもとに Cabocha が所定のルールを用いてモデルを作成する。これを用いて処理対象データにタグを付与する。
Step3　Step2 で付与されたタグ（抽出した医学的知識）が正しいかどうかを検証する。

実験結果

結果

- 文単位での正解率　42.4 %（抽出可能または主旨は抽出可能）
- 単語単位での正解率　67.5 %

文レベルで見た結果		単語レベルで見た結果	
処理の対象となる文の個数	262	処理の対象となる文に含まれる単語の総数	3,720
「抽出可能」であった文の個数	72 (27.5%)	単語としての「正解」の個数	2,511 (67.5%)
「主旨は抽出可能」であった文の個数	39 (14.9%)	単語としての「不正解」の個数	1,209 (32.5%)
「抽出不可」であった文の個数	151 (57.6%)		

図 3: 実験結果

実験結果

考察

・病院ごとのレポートの書き方の違いが影響
・SVM を適用する前に、処理対象データに使われている表現が学習データ中に出現するように言い換え処理を行うと、抽出精度が上がる可能性がある。

実験改善案 1

改善案 1

・同じ意味で、助詞や結語表現が異なる文章を多数機械的に作成し、学習用データに追加する。

・（言い換え例）

元の文: MRI にて脳に高信号域を認め、脳梗塞を疑う。

言い換え文: MRI にて脳に高信号域を認める。脳梗塞を疑う。

MRI にて脳に高信号域があり、脳梗塞を疑う。

実験改善案 1

結果

文単位での正解率は 51.9 %、単語単位での正解率は 73.2 %と精度の向上が見られた。

文のレベルで見た結果		単語レベルで見た結果	
処理の対象となる文の個数	262	処理の対象となる文に含まれる単語の総数	3,720
「抽出可能」であった文の個数	97 (37.0%)	単語としての「正解」の個数	2,724 (73.2%)
「主旨は抽出可能」であった文の個数	39 (14.9%)	単語としての「不正解」の個数	996 (26.8%)
「抽出不可」であった文の個数	126 (48.1%)		

図 4: 実験結果

実験改善案 2

改善案 2

・処理対象レポートの文を、意味が大きく変わらない範囲で表現が学習データとして使われている文章に近くなるように言い換えを行う。

・（言い換えの例）元の文:動脈癌は明らかではない。(抽出不可)
言い換え文:明らかな動脈癌はない。（抽出可能）

実験改善案 2

結果

・改善案 1 で抽出不可であった 126 文のうち 95 文が「抽出可能」、「主旨は抽出可能」となった。合わせると、正解率 88.2 %となった。

3〜4の段階で「抽出不可」であった文の個数	126
言い換えによって「抽出可能」になった個数	71 (56.4%)
言い換えによって「主旨は抽出可能」になった個数	24 (19.0%)
言い換えによっても「抽出可能」「主旨は抽出可能」に至らなかった	31 (24.6%)

図 5: 実験結果

まとめと課題

まとめ

・SVM を用いて読影レポートから「医学的知識」を抽出するための手法を示した。
・実際に学習用レポートとして作成したモデルを用いて処理対象レポートから「医学的知識」を抽出する実験とその結果を示した。
・SVM を用いた抽出の精度を向上させるための工夫として、両方のレポートに言い換え処理を行った結果、多くの文が正しく抽出された。

課題

・意味を大きく変えないような言い換えをどのようなルールで定めて機械的に処理するか、そのルールに従った結果、抽出精度がどの程度向上するのかを確認すること。