6Q−8

2種類のニューラルネットの縦列接続によるテキストの自動分類の試み

有田　英一
三菱電機㈱中央研究所

１．はじめに

近年、ニューラルネットワーク技術は音声認識や文字認識などのパタン認識、株価などの予測、ロボットなどの制御といった幅広い分野に適用され成果を上げつつある。知識情報処理の分野でもコネクショニストモデルに基づく自然言語理解システム[1][2]、情報検索システム[3]、知識ベースシステム[4]、などが提案されている。

本報告ではニューラルネットワーク技術の知識情報処理への応用の１つとして、テキストデータの自動分類について述べる。テキストデータの分類は通常は階層化されたキーワードで行なうが、①キーワードを付与するのが困難である、②新しい専門用語はキーワード体系にない、③キーワード間の関連が階層的なものしか設定されていないなどの問題点がある。一方、分類器としてのニューラルネットには汎化能力があり、テキストデータの分類にも適用できる可能性がある。もちろんニューラルネットの入力はそれ自身は意味を持たないパタンにしか過ぎないので高度な意味情報を含むテキストの分類に単純には適用できない。テキストを連続性のあるパタンに変換することが重要である。本報告では大規模な情報ベースに適用することを目標として、キーワードによらないテキストの自動分類手法の考察を行なう。その１つの手法としてテキストのコーディングを行なうコネクショニスト型ネットワークと、そのコーディングされたテキストの分類を行なう自己組織型の階層型ネットワークの縦列接続によるテキストの分類手法について考察する。

２．テキストの自動分類のためのニューラルネットに必要な機能

［１．追加学習が可能であること］
大規模な情報ベースを対象とする場合、全てのデータを対象として繰り返し学習させることは計算量の観点からみて非常に困難である。

［２．ネットワークの構造が可変であること］
大規模情報ベースを対象とする場合、予め出現する情報の構造を反映した入力構造を設定することが非常に困難である。

［３．蓄積される情報に従って分類が動的に変化できること］
技術の進歩に従って新しい専門用語が生まれるので、固定化されているキーワード体系では対処できない。そのためにネットワークは教師無し学習できることが必要である。

［４．自己組織化の方向がある程度制御可能であること］
教師無し学習によるパタン分類の場合、分類されるクラスの意味は前もって与えられない。しかしそれでは工学的に応用が困難であると同時に意味ある分類ができるとは限らない。人間の経験による高度な分類基準をネットワークの学習に生かすことが必要である。

３．テキストのコーディング

自然言語をニューラルネットワークで扱える表現に変換（コーディング）する方法として、マイクロフィーチャに基づいて概念を表現する手法があるが[1][6]、大規模な情報ベースを対象とする場合、マイクロフィーチャのセットを定めること及び概念をマイクロフィーチャ表現に自動的に変換することが困難である。本手法では各単語が入力の各ノードに対応するものとする。各ノードはその修飾関係の共起の頻度に応じた重みを持つリンクで結合する。（このようにして得られるネーットワークを単語ネットと呼ぶことにする。）このようなコーディングの考え方はニューラルネットのかな漢字変換への応用[7]でも用いられている。

あるテキストが単語ネットに入力されると、そのテキストに含まれる単語の修飾関係の共起関係の頻度情報を更新する。次にその単語に対応するノードの活性度を大きくし、それらの活性度をリンクの重みに比例して減衰させながら接続されているノードに伝播させる。このようにして得られた各ノードの活性度のパタンをそのテキストのニューラルネットに対する入力パタンとする。

Text Classification by Cascaded Two Types of Neural Networks
Hidekazu ARITA
Mitsubishi Electric Corp.

4．テキストの自動分類実験システムの構成

図1に示すように、単語ネットと2層のニューラルネットでテキスト自動分類システムを構成する。これは階層型ネットワークとコネクショニスト型ネットワークの縦列接続の形をしている。

単語ネット：
3．で述べた単語ネットである。共起関係によるノード間のリンクの重み付けを行なう。このネットワークにより生の形に近いテキストデータを連続性のある特徴量の空間に変換する。

入力層：
各ノードが単語に対応する。単語ネットのノードに対応する単語と1対1の関係で結びつけられる。

出力層：
入力情報の傾向を反映したある分類に従った結果を表わす。

学習：
入力層のノードと出力層のノードは完全結合されていて、最大値検出型仮説[5]に基づく自己組織化を行なう。学習データは情報 ベースに格納される全ての情報である。学習は情報が入力される度にインクリメンタルに行なう。入力される学習データの数が増すに従ってそのデータ全体の傾向を反映した分類となる。

コーディング：
入力されたテキストに含まれる単語ノードの活性度を大きくし、その活性度を接続されている単語ノードに減衰させながら伝播させる。

5．実験

3、4で述べたテキストの自動分類手法についてその効果を確かめるための予備実験を行なった。「（メーカ）が（新製品）を（述語）した」といった形の単純な単文のテキストを対象として自動分類を行なった。テキストデータを順次入力して行くにつれて分類がテキストの内容に応じて変化することが確かめられた。

6．おわりに

ニューラルネットの知識情報処理への応用の1つとしてテキストの自動分類の手法の提案を行なった。テキストのコーディングを行なうコネクショニスト型ネットワークとテキストの分類を行なう階層型ネットワークとの縦列接続の手法を提案した。

参考文献

[1] Waltz,D.L. and Pollack,J.B.: Massively Parallel Parsing: A Strongly Interactive Model of Natural Language Interpretation, Cognitive Science, Vol.9, No.1, pp51-74 (1985)
[2] 田村、安西：Connectionist Modelを用いた自然言語処理システム、情報処理学会論文誌、Vol.28, No.2, pp202-210 (1987)
[3] 田村、原、金子：Connectionist Modelによる「対話」としての情報検索システム、昭和62年度人工知能学会全国大会、pp159-162 (1987)
[4] 萩原：コネクショニストモデルを用いた知識ベースシステム、情報処理学会知識工学と人工知能研究回資料、89-AI-64-6 (1989)
[5] 福島：自己組織機能を持つ多層回路、電子通信学会論文誌、Vol.58-D,No.9,pp530-537,(1975)
[6] H.Ritter&T.Kohonen, Self-Organizing Semantic Maps, Biol. Cybern. 61, 241-254(1989)
[7] 鈴岡他：神経回路網の連想機能を用いたかな漢字変換方式、情報処理学会第40回全国大会、pp105-106 (1990)

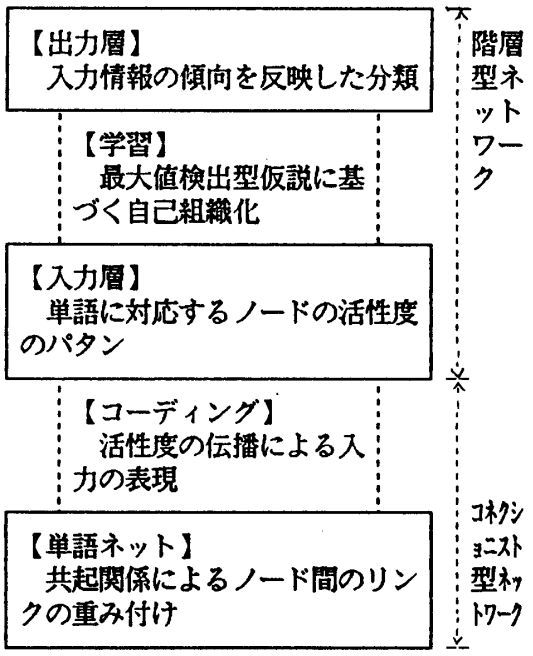

図1　テキスト自動分類実験システムの構成