Mathcadで学ぶ
機械学習の基礎と応用

連載　第12回　**文字データ（文字列）の抽出処理**

テクファ・ジャパン　香取　英男*
*かとり　ひでお：代表取締役　URL: http://www.tecpha.com/

機械学習においては、さまざまな物理現象から得られる離散的な点群データを取り扱うことが多い。すなわち、それらの点群データを種々の手法を用いて分析することによって、意味ある特性を獲得するのである。これまで数値データのみを取り上げて、その特性を把握する処理過程を調べてきた。ところで、機械学習で取り扱うデータは数値ばかりではない。下記に列挙するように、さまざまな形式のデータがある。

・テキスト（文字列）
・画像
・音声
・信号

前回は、テキストデータを処理するために、Mathcad Primeで用意されている組込み関数を紹介した。また、その使用例についても取り上げた。今回は、機械学習に応用するための処理を念頭に置いて、もう少しテキストデータの分析に近い処理を試みてみよう。

機械学習におけるテキストデータの分析の目的

テキストデータを処理する目的として、次のようなことがあげられよう。テキストデータを分析するということは、そのデータの内容から未知の情報を抽出することが主な目的である。これは、テキストマイニングと呼ばれる手法でもある。この処理を、手作業ではなく、コンピュータを介して機械学習で行うので、迅速に処理結果が得られる。また、その処理方法を繰り返して利用できるため、効率よく処理できることになる。

テキストマイニングの利用分野

ここでは、代表的な例をいくつかあげておく。

・顧客アンケートのニーズ抽出
・SNS、ブログ、掲示板などからの評判分析や情報抽出
・論文や特許などの技術文書からの技術マーケティング

テキストマイニングの具体的な処理内容

前項で示した適用領域において、その使用目的を果たすために、より具体的な処理例を示す。この中で、前回に紹介したMathcad Primeで用意されているテキスト処理用の組込み関数を適時組み合わせて用いる。

1. テキストデータの区切り文字に着目して箇条文を作成する処理例

本稿の冒頭の9行分のテキストデータを処理対象として例にあげてみよう。まず、“、”および“。”の2種類の区切り文字を用いて分割し、各節を抽出してみる（**図1**）。その際、区切り文字は削除する。ところで、Mathcad Primeでは、1つのかたまりのテキストを定義して表示すると、1行分として取り扱われる。対象にしたテキスト部の文字数は180以上あるので、この誌面では1行で表示しにくい。そこでいったん、5行に分けて定義した後、各行を1行に結合して処理対象のテキストデータを作成する。この分け方には特に意味はない。そのため、(1–1)から(1–6)までの本来必要のない余分な処理部が追加されている。