要 旨

ニューラルネットワークを用いた
テキストデータからの低次元特徴の抽出

川上 雄仁

従来，機械学習を用いたテキストマイニングの研究では，特徴量として単語ベクトルが広く用いられている．最も設計が容易な単語ベクトルは，One-Hot 表現 (1-of-K 表現) と呼ばれ，語彙数分の次元数のうち 1 単語に対する値のみが非零で，それ以外を 0 とするものである．One-Hot 表現では，ベクトルの各要素に単語が対応するため，次元が語彙数となり一般に数万次元になる．この One-Hot 表現を応用して，ある文書の特徴量を表すために，その文書で用いられる単語に対応する要素を 1（または出現回数や TF-IDF 値)，それ以外を 0 とするものがある．この特徴ベクトルを用いて機械学習を行うことを考える際に，次元が数万次元となるため「次元の呪い」の問題が発生する．このため，次元の削減を行う方法として様々なものが提案されているが，その一つに近年ニューラルネットワークの一種である Word2Vec を用いた単語ベクトルの低次元表現が提案されている．本研究の目的は，Word2Vec を用いた低次元特徴量と，One-Hot 表現を用いた高次元特徴量との識別精度の比較を行い，テキストマイニングにおける低次元特徴量の有効性を検証することである．本研究では，経済新聞記事からの株価予測，日本語メール文書からの SPAM メール判定，ある商品に対する英字レビュー文書の内容が商品についての肯定的か否定的かの判定を行う．各記事に対して，形態素解析処理を行い，日本語の名詞，動詞，形容詞を抽出する．レビュー文書のみ英語文書のため，不要となる記号等のみを削除する．訓練データに含まれる単語のみについて Word2Vec により低次元特徴を抽出する．得られた単語の特徴量を 1 記事中の単語分だけ合計することにより 1 記事の文書ベクトルとし，訓練データとテスト

データを作成する．One-Hot 表現については，単語の出現頻度を記事ごとに求め，単語ベクトルを作成する．教師ラベルとして，株価のデータセットについては，該当する記事の日の日経平均終値が 5%を越えて増加すれば上昇，5%を越えて減少すれば下降，値動きが ±5%以内であれば値動きなしとした 3 クラスとする．SPAM メールの判定については，SPAM か非 SPAM の 2 クラス，レビュー文書の判定については，肯定か否定の 2 クラスとする．結果として，Word2Vec による単語の低次元特徴量を用いたテキストデータ識別は SPAM メールのデータについては識別率が 95.2%となり，従来の単語表現より 5.2 ポイント向上することを示す.

キーワード　ニューラルネットワーク，Word2Vec，テキストマイニング，単語ベクトル

Abstract

Low-Dimensional Features using Neural Network for Text Data

Yuto KAWAKAMI

In the research area of the text mining using the machine learning, word vectors are widely used to represent text documents. Word vector expression is also called as one-hot expression (1-of-k expression), and only the value for a word is non-zero and the others are zero. Using one-hot expression, the number of dimension of the feature is over several thousands. The reason of the high dimension of word vector expression is that the dimension equals to the number of word used in all documents. The curse of dimensionality is caused by high dimension when machine learning is performed using feature vectors. In order to reduce the dimension, several methods have been proposed. Recently, low dimensional expression of word vector using Word2Vec has been proposed using neural network. The purpose of this research is to evaluate the low dimensional features using Word2Vec compared with high dimensional features using one-hot expression in text mining using machine learning. In this research, stock price prediction from newspaper, spam mail discrimination in Japanese, and reviews of shopping item in English are used to compare. First, words included in training data are input to Word2Vec and it outputs low dimensional features. Next all low dimensional word vectors in a news article are summed and the vector of the summation is treated as a feature vector of a news article. All articles are converted to article feature vectors, and they are divided into training and test data. For one-hot expression, frequency

of word appeared in an article are used to make a word vector. The training label of stock market prediction are three categories, price up, price down, and stationary price. The training label of spam mail in Japanese are two categories, spam and non-spam. The training label of reviews of shopping item in English are two categories, positive meaning and negative meaning. As a result, the accuracy using low dimensional features is higher than high dimensional features.

key words neural network, Word2Vec, text mining, word vector