要 旨

機械学習におけるランダム性を持つ特徴抽出法のテキストデータへの応用

藤森　夏輝

テキストデータが機械学習アルゴリズムで処理できるよう単語出現頻度などをもとに数値化された単語ベクトルは，本来は比例尺度のような定量的指標ではなく，順序尺度のような定性的な指標である．データマイニングに用いられる手法は特徴ベクトルの内積や距離が定義できる量的データの分類に効果のある数値計算で行うものが多いが，決定木を用いる Random Forest は，説明変数をランダムに選ぶことで様々な形の決定木群を構築するため，定性的なテキストデータに適した機械学習手法であると考える．説明変数をランダムに選ぶとき，従来は疑似乱数列を用いたランダムサンプリングが行われる．しかし，構築する決定木の深さの最大数，分類に用いる決定木の数，決定木の説明変数となる特徴の数が小さいと，生成される疑似乱数に偏りが発生してしまうことがある．この問題に対し，本研究では準乱数列生成器を適用する．準乱数列は一様の点列で構成されている．準乱数列を Random Forest へ適用することによって，木の深さの最大数，識別に用いる木の数，木の構築に用いる特徴の最大数がそれぞれ低く設定されても，ランダムサンプリングにおいて生成される数列が偏る可能性はなくなり，疑似乱数列を適用した同アルゴリズムが困難としていた，同条件下における高精度識別が可能となると考える．独自に収集した日本語 SPAM メールデータセット（データ数 : SPAM600，非 SPAM1000）を用いて，疑似乱数列と準乱数列を適用した場合において実験を行う．日本語 SPAM メールは，これまでの研究でサポートベクターマシンやニューラルネットワークに比較して，Random Forest が安定して高い識別精度となるという報告があり，決定木をベースとするためテキストデータとの相性が良いこ

とが考えられるため，識別精度比較の実験に用いる．その結果，木の深さが 2，構築に利用する特徴の最大数と識別に用いる木の数が 10 以下の条件の下で，変更後の識別率が 2〜3% 向上することを示す.

キーワード　Random Forest，疑似乱数列，Mersenne Twister，Low-Discrepancy 列，準乱数列

Abstract

Feature Extraction with Randomness for an Application to Machine Learning from Text Data

Natsuki Fujimori

In text processing, a word vector, which is converted from text document data, is usually used as a feature vector. A word vector is a histogram of word frequency or occurrence of a document. It is a numerical data, however it is not a quantitative data such as ratio scale. It is originally a qualitative data such as ordinal or nominal scale. Some methods for data mining using machine learning employ numeric calculation which can discriminate quantitative data that able to define the distance or inner product of feature vectors, but random forests employing decision trees is machine learning technique which is suitable for qualitative text data because it constructs various shapes of decision trees by choosing predictor values randomly. When choosing predictor values, the random sampling is performed using pseudo-random sequence. However, if the number of tree depth to construct, the number of decision trees to use discriminate, or the number of features predictor values of decision trees are small values, the bias may appear because of non-uniformness of pseudo-random numbers. In order to solve this problem, in this study, we propose an application of quasi-random number generator. quasi-random sequence generates uniform sequence of points. When the number of tree of depth, the number of trees using for discrimination, or the number of features constructing tree are small, the proposed method is able to achieve higher performance than pseudo-random. By using the original Japanese SPAM e-mail dataset (600: SPAM,

1000: non-SPAM), we perform the experiment of conventional and proposed methods. As a result, under the condition that the maximum number of tree is 2, and that the number of feature is under 10, the proposed method improves 2 or 3 percent of the precision.

key words Random Forest, Pseudo-Random Sequence, Mersenne Twister, Low-Discrepancy Sequence, Quasi-Random Sequence