N-004

データサイエンス記述問題からの各技術項目に対する理解度判定

Identifying Comprehension Fault from Word Occurrences in Writing Questions

安田健人† 島川博光† 原田史子†

Kento Yasuda, Hiromitsu Shimakawa, Humiko Harada

1. はじめに

IT 業界だけでなく，製造や物流，医療などの幅広い業界においてデータサイエンティストの需要が高まってきている．データサイエンスの技術習得には，確率統計，最適化，プログラミングなど多様な知識，スキルが求められ，挫折する者が多い．その要因として，何を理解していて何を理解していないかを知る方法がないことが考えられる．

本研究では自由記述の記載内容から理解できているかを割り出す手法を提案し，データ分析手法の理解度を推定する．行き詰まらず，理解している者の多くが知識の整理，順序立てた説明ができ，自分の言葉で表現することができることに本研究は着目する．記述問題で正解の文章に近い記述を回答した生徒は，学習した概念，手続きについても理解しているといえる．このような仮説に基づき，本研究は，自由記述を分析し，記述問題の正解への近さを判定する．

2. 埋め込みによる文書のベクトル化

2.1 次元圧縮と次元選択

高次元で表現された情報は，処理が複雑で可視化も難しい．そのため，必要な情報をできるだけ失わずに，次元を削減する必要がある．削減により，グラフ化が可能となり，標本の分布が直感的に理解しやすくなる．次元選択とは次元圧縮により次元削減が達成でき，それぞれに次元の意味を把握するために人間に理解しやすい次元を選択することである．適切な次元を選択することで正しく分類することが可能になる．

2.2 エンベティングとコサイン類似度

エンベディングとは，単語や文などの自然言語の情報を，それらの意味を表現するベクトル空間に配置することをさす．単語や文の間の類似度の計算のためには，単語や文をベクトル化することが一般的である．コサイン類似度は，2 つのベクトルの内積で算出される，一方の働きが，他方の働きにどれほど貢献するかを表す指標である．

2.3 文書のベクトル化

専門分野をテキストで説明した記述には，より基礎的な複数の分野の概念や手続きが含まれている．与えられた新しい記述が正しいものかを判定するのは，基礎分野ごとに正しい概念や手続きが使われていることを確認することになる．記述を分析するために，各基礎分野に次元を割り当てると，テキスト記述は高次元空間で表現される．

† 立命館大学，Ritsumeikan University

3. 記述問題の正解への近さの判定

3.1 理解不足の概念と手続きの推定

本論文では,学習者の理解を確認するために出題した記述問題において,学習者の解答が正解に近いかを判定する手法を提案する.さらに,本研究では,理解が得られていない学習者を適切に指導するために,専門分野における,どの概念,手続きが理解できていないかを,記述問題に対する解答をテキスト分析することで判別する.提案手法の概要図を図 1 に示す.

図 1：手法概要図

3.2 正解と解答のエンベティング

記述問題における正解例や学習者の解答は高次元の情報を含んでいるテキストデータである．Doc2Vec[1]を使用して，正解例と解答のテキストを低次元データにエンベディングする.次元を圧縮することですべての情報を加味できる。

高次元では各次元の意味を把握しにくい．理解しやすい次元を選択する必要がある.学習者が作成した記述に出現する単語を，Doc2Vec を使用して可視化する.文書ベクトルから各テキストの関係性を示すために,高次元ベクトルを 2 次元に圧縮する．

3.3 次元ごとの重要単語の算出

重要単語の抽出では,ナイーブベイズ[2]を利用する.文を単語に分割するために MeCab[3]により形態素解析する.その結果を使って生成する単語文書行列から多項のナイーブ・ベイズ・モデルを作成

し,特定分野に属する文書に出現する単語の出現確率を算出する．確率が高い順に単語を列挙すれば,その分野に属する文書での重要単語が得られる．

3.4 理解不足の概念の抽出

ロジスティック回帰モデルにより正解者と不正解者で使用された複数の重要単語の中から特に正解と不正解を分けるうえで重要な単語の重みを求める.本研究は，特に重要な単語の重みを求めることで理解不足の概念の抽出をすることができると考える.単語の重要度を算出することで，どの概念と手続きが，正解,不正解を分類するうえで大きく影響を与えるかを発見することができる．

4. 実験及び実験結果と考察

4.1 実験概要

本研究では理解度検証問題,確認テストに注目する．理解検証問題，確認テストでは，指定した 2 つの機械学習アルゴリズムが属する，次元圧縮という分野での概念と手続きについての問を出題した.40 点満点の理解検証問題では，正誤問題 16 問と 30 文字以上で説明する記述問題 4 問を解いてもらう．確認テストでは，100 文字以上の記述問題 2 問を出題した．

4.2 Doc2Vec による類似度の推定

Doc2Vec による文章の分散表現化の精度を確認する．確認テストの 2 問で学習者が記述したテキストと，正解例のテキストで類似度を算出する．Doc2Vec を用いて，それぞれのテキストを表すベクトルを得る．コサイン類似度により，学習者のテキストが正解のテキストにどれだけ近いかを確認する．

実験 2 で実施した確認テストの記述問題 2 問を,正解者と不正解者に出現する単語の類似度について Doc2Vec を使用して分散表現に変換して可視化する．可視化することで正解者と不正解者を分類できるか確認する．図 2 で，問題ごとにコーパスを用意すれば，正解者と不正解者のそれぞれがよく使用する単語を直線で分類できることを示す．

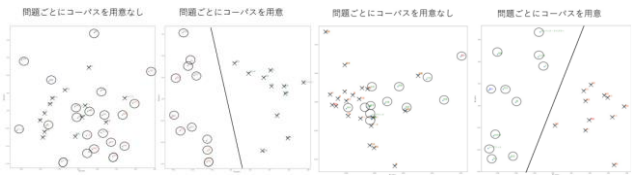

図 2：確認テスト 1， 2 での単語の分散表現

4.3 ナイーブベイズによる重要単語の抽出

出現確率が 0.02 を超える単語を重要単語とする．正解とのコサイン類似度が高い生徒は理解が高く，コサイン類似度が低い生徒は理解ができていない．図 3 の確認テスト 1 と確認テスト 2 より，正解とのコサイン類似度が高い生徒は多くの重要単語を使用しているが，コサイン類似度が低い生徒は重要単語をあまり使用していないことが読み取れる．

図 3：確認テスト 1， 2 での重要単語の抽出

4.4 ロジスティック回帰による重要単語の算出

図 4 で，丸で囲んだ単語は，ロジスティック回帰による正解と不正解の分類モデルで重みの絶対値が大きい単語である．これらは，正解と不正解を分類するうえで重要な単語である．図 4 より，確認テスト 1 の重要単語は NMF の特徴,仕組みに関わるものであることがわかる．

図 4：確認テスト 1， 2 での重要度

5. おわりに

実験の結果，教科書上の記述をコーパスとすることにより，本手法は，正しい理解を得ている学習者を判別できることがわかった．それだけでなく，概念を理解できていない学習者が使う単語は正解例が使う単語とほとんど重なりがないことがわかった．正解例に現れるが，学習者の回答には現れない単語を調べることによって，その学習者が理解できていない概念を特定することができる．

参考文献

[1] Quoc V． Le， Tomas Mikolov: “Distributed Representations of Sentences and Documents”， ICML’14， Vol. 32， pp. 1188-1196， 2014.
[2] Andrew McCallum, Kamal Nigam, “A Comparison of Event Models for Naïve Bayes Text Classification”, AAAI Conference on Artificial,19
[3] 工藤拓，山本薫，松本裕司：“Conditional Random Fields を用いた日本語形態素解析”， 情報処理学会研究報告．自然言語処理研究会報告 161，pp．89-96，2004.