XXXXXXXX 若手社員 A

VAE を用いた遺伝子発現プロファイルデータの次元削減とクラスタリングへの応用

1. プロジェクトの概要

近年になり、シングルセル解析という一細胞単位での遺伝子発現解析ができるようになりました。(従来はバルク解析)

これにより、一細胞単位での各細胞の正確な遺伝子発現量を行列データとして取得できるようになりました。

本データには多くのノイズが含まれており、非線形的に柔軟な次元削減を行いたいというモチベーションから、

変分自己符号化器(VAE)の潜在空間を用いた次元削減の研究を行いました。

2. プロジェクトにおけるあなたの立場

研究当事者です。大学院の教授との議論を交えながら、データ分析の全てを一人で行いました。

また研究の問題設定も、准教授と会話しながらテーマを設定しました。

3. 直面した技術的な課題や困難に対する創意工夫

本研究の対象である、遺伝子発現プロファイル は細胞種を複数個含むため、潜在空間の事前確率分布においてモードを細胞種の数だけ持つような柔軟な表現をすることが重要です。しかし、従来の潜在空間の事前確率分布である正規分布では 1 つの正規分布でしか表現できませんでした。そこで、Meta-GMVAE: Mixture of Gaussian VAE for Unsupervised Meta-Learning (Dong Bok Lee 著) を参考に混合ガウスモデルを事前確率分布に用いることで、各正規分布が各細胞種(クラスタ)を示すようにし 、柔軟な表現を達成することができました。これにより、モデルの精度の向上とともに、潜在空間におけるクラスタの分布同士を分離することができました。そして、潜在空間に対して k-means、DBSCAN、GMM をかけることでクラスタリングへも応用しました。

4. あなたが解決した・または解決できなかった要因の考察

入力データをうまく潜在空間で表現することができたため、入力データの無駄な情報(ノイズ)を減らすことができました。かつ、潜在空間におけるクラスタリングを実現した。 一方で、正解のクラス数(細胞種数)を恣意的に設定したことが課題として残っています。これについては、潜在空間を混合ガウス分布にしていることを踏まえ、EM アルゴリズム等を用いることで、おおよそのクラス数を算出することができたと考えられます。

5. 将来、あなたが取り組んだプロジェクトが、何にどのように役立つか

遺伝子データのより高精度な分析による 病気の早期発見です。遺伝子から、その人の罹患しやすい病気を未然に判別する技術に役立ちます。

また、異常検知や疑似データの生成など、他分野におけるデータ解析に技術転用することが可能です。

XXXXXXXX 若手社員 B

タクシープローブデータによる道路交通異常の自動検出

1. プロジェクトの概要

「積雪という特殊状況下で、通行止めなどの道路交通障害が起きている地域を異常地域として自動検出する」という異常検知に関するテーマで研究を行いました。　道路交通障害の検知にはウェブカメラなどの映像データを通常必要とするケースが多いと考えられますが、本研究ではタクシーのプローブデータを用いて、タクシーの速度の変化に着目することで異常を検出するというアプローチで取り組みました。

2. プロジェクトにおけるあなたの立場

研究課題の設定、データ整形、モデルの作成、関係者への説明を単独で行いました。

3. 直面した技術的な課題や困難に対する創意工夫

分析に際して、「道路交通障害が起きている地域のタクシーの平均速度は、その地域の普段(道路交通障害が起きていない時)の平均速度と比べて遅くなる」という仮説を立て、2 つの平均速度の比較による異常の検出を試みましたが、積雪などの東京全土に影響が及ぶイベント時は、「道路交通障害の有無に関わらず、ほとんど全ての地域で速度の低下が検出されてしまい、道路交通障害のある地域を正しく検出できない」という課題に直面しました。 その際に、積雪の日だけでなく、積雪のない日の地域別速度を可視化したり、平均速度を算出するなど、積雪の日のデータに捕われるのではなく大局的にデータを観察するなどの試行錯誤の末、「東京全域の平均速度と各地域の平均速度の差」を新たな指標として算出し、この差を積雪時と通常時(積雪ではない時)とで比較するという工夫により、積雪による影響を適切に除去した上での分析を可能にしました。

4. あなたが解決した・または解決できなかった要因の考察

今回分析に使用したデータからは速度低下の原因を明確化することが出来ないため、厳密な精度検証（本当に障害が起きていたか）を行うことが難しいという課題が残りました。例えば、高速道路の事故渋滞情報や Twitter の呟き情報などから、いつ・どこで・どんな交通障害が起きたかというデータを網羅的に収集することが出来れば精度検証が可能になると考えます。

5. 将来、あなたが取り組んだプロジェクトが、何にどのように役立つか

通行止めや立ち往生、渋滞といった道路交通障害は現状、交通監視カメラやパトロールによる発見が主であり、全ての道路交通障害を迅速に把握する事は困難であるため、機械学習技術を用いた異常検知の自動化は、業務負担の軽減という観点で役立つと考えています。

XXXXXXXX 若手社員 C

季節を考慮した俳句の自動生成システムにおける評価機構に関する研究

1. プロジェクトの概要

私が取り組んだ研究は、俳句の自動生成及び自動生成された俳句に対する評価機構の開発を通して、定性的な指標である俳句の”良さ”に影響を与えている要因の分析です。俳句の作成及び評価時の情報処理プロセスについて、心理学的アプローチによる研究は行われてきているものの、システムを作って動かすことによる理解を目指す「構成論的アプローチ」による研究はあまり行われていません。本研究は構成論的アプローチに基づき心理学的知見を元に計算機による俳句の自動生成システムの開発と、計算機による生成された俳句に対する評価機構を検討しました。

2. プロジェクトにおけるあなたの立場

新規研究テーマなので、指導教員とコミュニケーションをとりつつ主体的に推進するような立場でした。
具体的には以下の活動に取り組みました。
・研究テーマの立案:学部の研究テーマの課題を整理し、発展する形かつ修士で取り組める程度の粒度感でどのような方向性が適切か論文サーベイし指導教官に提案を行いました。
・実験・分析:俳句評価の為のデータを集めるために実験デザインを設計し、クラウドソーシングで自動生成された俳句の評価実験・分析を行いました。

3. 直面した技術的な課題や困難に対する創意工夫

俳句の評価に関して、季語がとても重要であるという研究結果が多数あり、季語に関する特徴量設計を行いたいと考えていました。しかし、季語は明確に決まってなく、辞書の作成が難しく、また季語の中に強弱があり、適切な季語の推定タスクがかなり難しいことがわかりました。この課題に対して、単語の分散表現技術を活用し、各季節と単語の類似度を計算することで対象の句における季語の推定することで俳句評価モデルの重要な特徴量の設計を行うことができました。

4. あなたが解決した・または解決できなかった要因の考察

自動生成された俳句の評価指標に関して示唆することができたものの、評価モデルの精度があまり高くない結果となりました。これは、評価モデルの学習に使うデータを一般人の評価データを用いたからだと考えられます。一般人のデータにおいて、俳句の慣れ親しみの度合いで群を分けモデル構築を行い、比較をしたところ、慣れ親しんだ人たちの評価に一貫性があることとが示唆されました。このことも踏まえ、俳句のような芸術鑑賞はその対象への理解度が評価に強く影響を与えるため、プロの俳人の評価データを構築し、俳句評価モデルを学習する必要があります。

5. 将来、あなたが取り組んだプロジェクトが、何にどのように役立つか

近年、文章や画像などコンピューターによるコンテンツ自動生成が行われているが、多くが言語や画像としての自然さのみに注目していることが多いです。これに対し、人間の認知処理に基づいて評価指標を設計し、評価機構を作成することで、流暢さと感性的な評価の 2 軸を考慮した自動生成を行えると考えます。特に、本研究は俳句を対象としているが、言語モデルや語の分散表現を用いて、特徴となる語の類似度や流暢度評価指標として用いるアプローチは、他の言語的な作品の評価に応用できると思います。これにより、将来、人とコンピュータが共創して素人でも気軽にコンテンツの作成・理解ができるような社会の実現に向けて貢献できると考えています。