学会記事

Evaluating Evaluation Measures for Ordinal Classification and Ordinal Quantification

酒井 哲也†

1 なぜ ACL-IJCNLP 2021？

私は普段 ACM SIGIR（情報検索の Special Interest Group）のコミュニティで活動しており、筆頭（というより単著）で ACL に論文を投稿したのは今回が初めてではないかと思います。なお IJCNLP には 2005 年（第 2 回 IJCNLP）にやはり評価指標に関する論文を投稿し、バッサリ斬られた記憶があります。今回 ACL 2021 に投稿した理由は 3 つあります。 (1) SIGIR 2021 のプログラム委員長を務めており、自分が筆頭の論文を会議に投稿するのは避けたかったため。 (2) ACL 2020 で ordinal classification タスク（後述）のための評価指標を提案した論文が発表され、ACL でこのような論文が受け入れられる余地があると悟ったため。 (3) 本研究で扱った ordinal classification および ordinal quantification タスク（後述）は、情報検索タスクというよりも、SemEval†や対話関係のタスクなど、自然言語処理の範疇であるものが多いため。

幸い、今回は無事論文が採録されました。この経験を糧に、今後は学生らとともにコミュニティを横断した研究を進められればいいなと思っております。

2 なぜ「評価指標の評価」？

私は「評価指標の評価」に関する研究を沢山やってきました。我ながら地味な研究だからあとは思います。評価指標の設計および選択は、研究の根幹をなす「メタな研究」であり、大変重要です。あるタスクにおいて、システムがどれくらい望ましい挙動を示しているか、システム A とシステム B ではどちらがより望ましいか、それを決めるのが評価指標だからです。もし不適切な評価指標（例えば我々が目指すシステムを正当に評価できないもの）がコミュニティで用いられていると、コミュニティの研究全体がおかしな方向に行ってしまう可能性もあります。繰り返しますが、評価指標に関する研究は大変重要です。

以下、本題に入ります。本研究では、ordinal class （順序的クラス）を扱う 2 つの形態のタス

† 早稲田大学
† https://semeval.github.io/
クを扱っています。伝統的な文書分類タスクでは、与えられた記事を「政治」「社会」「スポーツ」などのクラス（カテゴリ）に分類することなどを考えますが、このときのクラスは nominal（名義的）です。すなわち、クラス間の順序という概念がありません。これに対し、例えば与えられたツイートを「非常にポジティブ」「ポジティブ」「ニュートラル」「ネガティブ」「非常にネガティブ」に分類するタスクは、ordinal classification タスクであるといえます。

なぜ ordinal class を扱う場合に特別な評価指標が必要なのでしょうか。名義的クラスを扱う従来の評価指標をそのまま流用することは適切でしょうか。答えは NO です。何故 NO であるかは、上記のツイート分類タスクにおいて、あるツイートの正解ラベルが「非常にポジティブ」であるにも関わらず「非常にネガティブ」と分類してしまう「大外し」システム A と、「ポジティブ」と分類してしまう「おいしい」システム B を考えれば明らかです。名義的クラスを扱う評価指標では、システム A もシステム B も共に「不正解」であり、システム B の挙動のほうがより好ましいという事実が数値化できません。これでは ordinal classification を行うシステムをうまくチューニングすることも、正当に評価することも難しいでしょう。

3 Ordinal Classification タスクに適した評価指標

図 1 を用いて ordinal classification（以下 OC）のタスク仕様を説明します。この例は SemEval-2017 Task 4 Subtask C (Rosenthal et al. 2017) などを参考に作りました。OC タスクでは、n 個の評価用トピックを用いてシステムを評価します。このうちのひとつが図中に示した“Ariana Grande”（米国の歌手名）であると考えてください。システムには、各トピックに付随する N 個の事例（ここではツイート）が与えられます。一般に、N の値はトピック毎に異なります。図中では、簡単のために、1 番目、2 番目、および N 番目のツイートのみについて、それぞれがある特定のシステムによりどのクラスに分類されたか、また各ツイートの真のクラスが何で

![Test topic: “Ariana Grande”](image)

"I worship @ArianaGrande"
Tweet 2: "Who is Ariana Grande?"
Tweet N: "I love @ArianaGrande"

Gold classes

| Highly Pos | Pos | Neu | Neg | Highly Neg |
|------------|-----|-----|-----|------------|

PERFECT! 1
CLOSED! 2

Numbers in the cells add up to N

図 1 Ordinal classification タスクのテストトピックの例とその混同行列の一部.
あるか、混同行列により示しています。具体的には、2 番目のツイートの真のクラスは Neutral であり、システムがこれを正しく Neutral に分類できたことが対角線の「1」により示されています。一方、1 番目と N 番目のツイートの真のクラスは Highly Positive ですが、このシステムはこれらを Positive に分類しています。お話しです。この図からわかるように、OC タスクでは、各事例が混同行列の対角線上に載るか否かではなく、各事例が対角線からどれだけ離れたセルに載るかを考慮して評価する必要があります。なお、各トピックに対応する混同行列より算出された評価値は最终的に n 個得られるので、通常は評価値の平均によりシステムの優劣が議論されます。

上記の性質を満たす評価指標に、Weighted κ、Krippendorff’s α、SemEval で用いられた MAE^M および MAE^p、ACL2020 で提案された CEM_{ORD} (Amigó et al. 2020) があります。ただし、このうち MAE^p は class imbalance （正解が特定のクラスに大きく偏った状況）をうまく扱えません。これに対し、その他の指標は、正解の分布の偏りによらず全てのクラスの重要性を同等と見なしして評価を行うことができます。

私の論文では、上記のものに限らず、混同行列から計算できる評価指標を以下の 3 つの観点から比較しました。第一に、全ての評価指標の対についてシステム値位の順位相関をもとめ、似た挙動を示す評価指標を調べました。第二に、評価用トピック集合をランダムに 2 分割することを繰り返し、各評価指標のシステム値位がトピックの顕れにどれだけ依存するか、すなわちデータの選定に対する頑健性を調べました。第三に、各評価指標について、判別能力 (discriminative power) を調べました。これは、全システム対の統計的検定を行った場合の p-value の系列を求めることがないように、各指標が実験から結論を導くのにどれほど役に立つかを調べるもので (酒井 2015)。検定にはランダム化 Tukey HSD 検定 (Sakai 2018b) を用いました。

SemEval の 2 つのデータセットを用いた実験結果をもとに、私の論文では OC タスクに適した評価指標として Weighted κ と Krippendorff’s α を推奨しています。ただし、κ には弱点があります。それは、SemEval のように「常に Highly Positive に分類するシステム」や「常に Neutral に分類するシステム」のような複数のベースラインがある場合、κ はこれらのいずれに対しても全てのトピックについて 0 点を与えないという点です。この性質は κ の定義式から簡単に証明できます。SemEval では MAE^M が OC タスクの評価に用いられてきましたが、私の実験では、この指標は頑健性と判別能力の観点から κ や α に及びませんでした。ACL2020 にて提案された CEM_{ORD} についても同様でした。

4 Ordinal Quantification タスクに適した評価指標

図 2 を用いて ordinal quantification (以下 OQ) のタスク仕様を説明します。この例は SemEval-2017 Task 4 Subtask E (Rosenthal et al. 2017) などを参考に作りました。評価用トピックが n
図 2 Ordinal quantification タスクのテストトピックの例と正解分布・推定分布。

個々の各トピックに $N$ 個の事例（ここではツイート。一般に $N$ はトピック毎に異なる。）がある。各事例に真のクラスが付与されているところまでは OC タスクと同じです。こちらもやはり $n$ 個の評価値の平均によりシステムに優劣を評価します。違いは、図中の gold distribution（真の分布）で示したように、個々の事例がどのクラスに属するかは問題にせず、事例がクラス上にどのように分布しているか、すなわち各クラスにいくつの事例が属するかののみを問題にする点です。すなわち、OQ タスクを解こうとするシステムには、この真の分布を推定することが求められます。このような形態のタスクとしては、上記 SemEval タスクの他にも対話破綻検出チャレンジ（雑談対話中のシステムの各発話が「対話破綻」「対話破綻の可能性あり」「対話破綻なし」のいずれであるかに関して、複数の判定者が付与したラベルの分布を推定する）(角森 他 2020) や NTCIR の Dialogue Evaluation タスク（ヘルプデスク対話に対し複数の判定者が付与した 5 段階の対話品質スコアの分布を推定する）(Zeng et al. 2020) などがあります。個人的には、複数の判定者のラベルを多数決などによりひとつに集約してしまうのではなく、個々の判定者の観点を真の分布という形で保持した上でシステム評価を行う OQ タスクのようなアプローチは、今後重要性が増すと考えています。

さて、OQ タスクの評価には真の分布と推定分布との比較が必要となります。よく知られた Mean Absolute Error (平均絶対誤差) や Jensen-Shannon Divergence などの評価指標をこの場合に用いることは適切でしょうか。答えは NO です。これらの指標は、真の分布と推定分布のずれを各クラスについて求め、これらを単純に足し合わせるものだからです。図 2 の真の分布は Highly Positive に偏っていますが、推定分布（実線部分）は逆に Highly Negative に偏ってしまっているという意味で「大外し」です。上記のような評価指標は ordinal class を考慮できないので、「大外し」であることが数値化できません。図中に点線で示したように、推定分布の Highly Negative のクラスに載っている 5 個のブロック（すなわちツイート）のうち、4 個を Neutral のクラスに移動した別の推定分布を考えてみてください。こちらのほうがだいぶ真の
Evaluating Evaluation Measures for Ordinal Classification and Ordinal Quantification

分布に近いですが、ordinal class を考慮しない指標ではこの 2 つの推定分布の違いが数値化できないのが明らかです。

OQ タスクを正当に評価できる評価指標としては、NMD（Normalised Match Distance、Earth Mover’s Distance と本質的に同じ）と、RNODE（Root Normalised Order-aware Divergence）（Sakai 2018a）およびこれに対称性をもたせた RSNODE があります。NMD に相当する Earth Mover’s Distance は SemEval の OQ タスクで、RSNODE は NTCIR の Dialogue Evaluation タスクで用いられています。私の論文では、OQ タスクの評価には RNOD と NMD を併用し、前者を主要な指標とすることを推奨しています。前述のデータの選定に対する頑健性の観点および判別能力の観点から、RNOD のほうが NMD よりも総合的にはよい結果を示したためです。

5 まとめ

本研究に関する合計約 90 分のチュートリアルビデオ、99 枚のスライド、および ACL2021 向けに提出した 12 分間のビデオなどを公開しています。2 ご興味があればご覧ください。

参考文献

Amigó, E., Gonzalo, J., Mizzaro, S., and de Albornoz, J. C. (2020). “An Effectiveness Metric for Ordinal Classification: Formal Properties and Experimental Results.” In Proceedings of ACL 2020, pp. 3938–3949.

Rosenthal, S., Farra, N., and Nakov, P. (2017). “SemEval-2017 Task 4: Sentiment Analysis in Twitter.” In Proceedings of the 11th International Workshop on Semantic Evaluation, SemEval ’17, pp. 502–518, Vancouver, Canada. Association for Computational Linguistics.

酒井哲也 (2015). 情報アクセス評価方法論: 検索エンジンの進歩のために. コロナ社. [T. Sakai (2015). Information Access Evaluation Methodology: For the Progress of Search Engines, Corona Publishing.].

Sakai, T. (2018a). “Comparing Two Binned Probability Distributions for Information Access Evaluation.” In Proceedings of ACM SIGIR 2018, pp. 1073–1076.

Sakai, T. (2018b). Laboratory Experiments in Information Retrieval: Sample Sizes, Effect Sizes, and Statistical Power. Springer.

角森唯子, 東中竜一郎, 高橋哲朗, 稲葉通将 (2020). 対話破綻検出チャレンジ 3 における対話破綻検出の評価尺度の選定. 人工知能学会論文誌, 35 (1). [Tsunomori et al. (2020).

https://waseda.box.com/acl2021sakai-videos-and-slides
Selection of Evaluation Metrics for Dialogue Breakdown Detection in Dialogue Breakdown Detection Challenge 3. Transactions of the Japanese Society for Artificial Intelligence, 35 (1), DSI-G_1-10].

Zeng, Z., Kato, S., Sakai, T., and Kang, I. (2020). “Overview of the NTCIR-15 Dialogue Evaluation Task (DialEval-1).” In Proceedings of NTCIR-15, pp. 13–34.

略歴

酒井 哲也：早稲田大学基幹理工学部情報理工学科教授．Naver Corporation 研究統括顧問．国立情報学研究所客員教授．ACM Distinguished Member．ACM SIGIR Vice-chair (2019–2022). http://sakailab.com/tetsuya/