

運転時系列データへの簡便なアノテーション手法の提案

Simple Annotation Method for Time-Series Driving Behavior Data

田中雄介† 栗原慶典† 那和一成† 谷口忠大†† 高野敏明††

竹中一仁†† 坂東誉司†

Yusuke Tanaka Keisuke Kurihara Kazunari Nawa Tadahiro Taniguchi Toshiaki Takano
Kazuhiro Takenaka Takashi Bando

1. はじめに

近年の情報技術の発展により、自動車や携帯端末などの移動体から大量のストリームデータが収集できる環境が整いつつある。これらの大量のストリームデータをデータ提供者に関する Web 上のデータなどと合わせてデータセンターに集積し、自動車のドライバや携帯端末利用者への種々の支援を提供する Cyber Physical System の構築が期待できる。これにより、従来よりも利用者と協調できる安全運転支援システムや情報提供サービスの実現が期待できる[1](図1)。

このようなシステムを構築・運用するためには、集積した大量のストリームデータを解析する必要があり、一部または全部のデータに対して、データの性質を示すアノテーションを加える作業が生じる。しかしながら、ストリームデータに対するアノテーションには以下に示す二つの大きな課題がある。

一つは、アノテーションされたラベルに作業者ごとの解釈のばらつきが含まれることである。特に「走りやすさ」や「混雑度」といった主観的なラベルの付与基準を明確に規定することは容易ではない。また、ストリームデータにおける特定のパターンを表現するラベルを付与するタスクでは、ラベルの始点・終点にもばらつきが生じる。更に、入力データや解析器の内部モデルが持つ特定パターンの始点・終点と、作業者がアノテーションを加える時に観測する当該パターンの始点・終点が一致しない場合も存在する。自動車の運転時系列データの例では、先行車や交差点という具体的な対象物でも、作業者によってラベルを付与する範囲は変わる可能性が存在する。

もう一つは、アノテーション作業自体の煩雑さである。これは、現在普及しているアノテーション方式が移動体によって収集される大量のストリームデータの処理に適していないことによる。映像解析や音声解析では、数サンプル/s～数十サンプル/s に区切られたデータの各サンプルに対してラベルを付与する方式のアノテーションツールが広く知られている[2,3]。この方式は少数のデータに対して良く吟味された詳細なラベルを付与する用途には優れているが、大量のデータへの簡便なアノテーションには不向きと考える。

本稿では、意味単位で分節化された運転時系列データに対するアノテーション手法を提案する。また、運転時

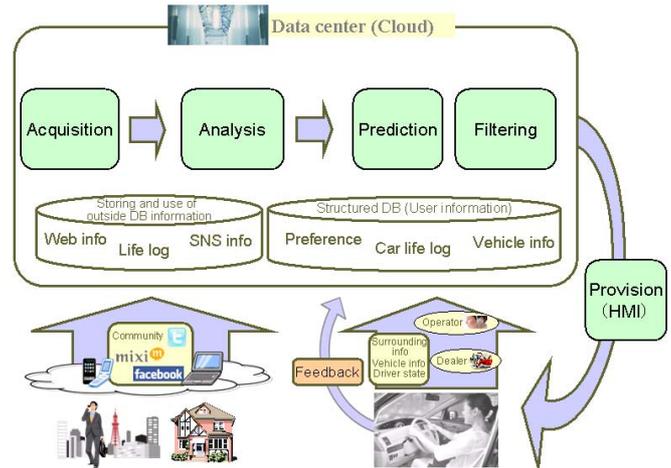


図1. 自動車の運転支援のための Cyber Physical System

系列データの二重分節解析結果に対するアノテーションが、運転時系列データの活用につながることを示す。

2. 関連研究

武田らは、運転時系列データに逐次アノテーションを加えながら収集する取り組みを行っている[4]。この取り組みでは、前述の課題をデータ収集環境および作業者の育成も含めたアノテーション作業環境を統一することで解決している。しかし、このような統一された環境を大規模に整備するためには多大なコストが生じるため、広く一般車両からのデータ収集としての利用は難しい。

これに対し、クラウドソーシングは大量の情報を処理するための有用なアプローチである。しかしながら、一般にクラウドソーシングを用いて収集されたデータは、良く整備された作業環境で収集されたデータに比べて精度が劣る。特にアノテーションにおける作業者間のばらつきを抑えることは困難といえる。クラウドソーシングを用いて収集したアノテーションを活用する方法もいくつか提案されている [5]。これらの多くは言語処理や画像処理などの、データの分節点が比較的明確な対象を扱っており、分節点が明確でないストリームデータへのアノテーションに単純に適用するのは難しい。また、実際にクラウドソーシングを使ってデータのアノテーションを行うためには、使いやすく分かりやすいアノテーションツールに代表される適切な作業環境を提供することが重要となる。

† (株) トヨタ IT 開発センター

†† (株) デンソー

††† 立命館大学



図 2. ラベル付けツール操作画面

ラベルの始点・終点の曖昧性を低減する取り組みとして、音声言語処理の分野では、人手で音声信号をテキストに書き起こし、形態素解析ツールを用いて分節した結果に対してアノテーションを行う手法が以前から知られている[6]。従来のアプローチが時間単位の区切りに対するアノテーションであったのに対し、[6]で用いられているのは意味単位の区切りに対するアノテーション手法であり、これによって、アノテーション区間の始点・終点のばらつきが抑えられている。さらに、時間単位で区切られた信号へのアノテーションに比べて分節の数を抑えられるため、作業者の負担も低減できる。

自動車の運転時系列データに対する分節化手法として、谷口らの手法がある[7]。谷口らの手法では、運転時系列データに二重分節性を仮定して、運転時系列データ中の比較的長い連続な運転行動を一連のチャンクとして抽出している。また、当該手法に依れば良く似た運転行動を同じチャンク(頑強チャンク)として抽出することも示されている[8]。これは、運転時系列データにおける意味単位の分節といえる。

3. 提案手法

運転時系列データへのアノテーションを困難にしている大きな要因として、ラベルの始点と終点が曖昧なことは既に述べた。これを解決するために、本稿では、谷口らの提案する二重分節解析によって分節した各区間にラ

ベルを付けることを基準とするアノテーションツールを作成した(図2)。

作成したツールは、データ取得時の走行経路全体と分節点を示すアイコン(図中の旗のアイコン)、データ取得時に撮影した動画像と連動して当該時点の走行位置を示すアイコン(図中の車のアイコン)を表示する機能を備える。これにより、位置情報を伴うストリームデータを扱う場合に従来の汎用的なアノテーションツールに比べて多くの情報を作業者が得ることができる。加えて、位置情報と時間情報の両方をインデクスとして区間を選択できるため、従来のアノテーションツールに比べてラベルを付与したい区間を直観的に選択することができる。

4. 評価実験

4. 1. 実験内容

実際に分節化された運転時系列データに対して試作したツールでラベルを付与し、評価を行った。

本稿の評価では運転時系列データの二重分節解析結果[7]を用いているため、二重分節解析器について簡単に紹介する。二重分節解析結果の概要を図3に示す。運転時系列データに対する二重分節解析では、まず運転時系列データの inputs を音声言語処理における音素に当たるセグメントに分割し、セグメントの連なりを単語に該当する

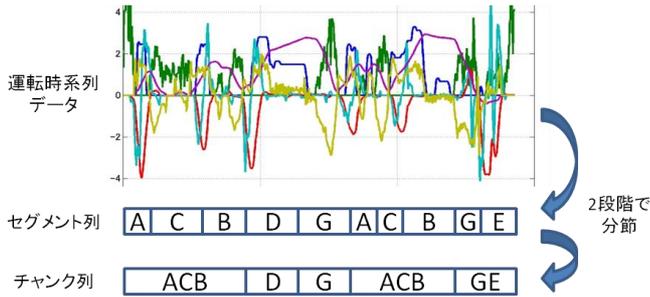


図 3. 二重分節解析の概要

表 1. 実験で用いたラベルセット

ラベル名	付与基準
走りやすい	自分が運転している場合を想定、主観的で良い
走りにくい	同上
狭い	実際の道幅だけでなく、受ける印象として
広い	同上
1車線	対向・一方通行は問わない1車線の道路
片側1車線	片側1車線の道路
片側2車線以上	片側2車線以上の道路
区間に交差点を含む	交差点から影響を受けていると思ったら

表 2. データ諸元

項目	内容
走行データ	<ul style="list-style-type: none"> 東京都港区赤坂近辺を走行した同一コース3周分の走行データ 各周回を dataset A,B,C とする 1周の走行距離は約 2.5km 1周の走行時間は約 10~12分
ラベルデータ	5名の作業者が表1の基準で付与したラベルデータ

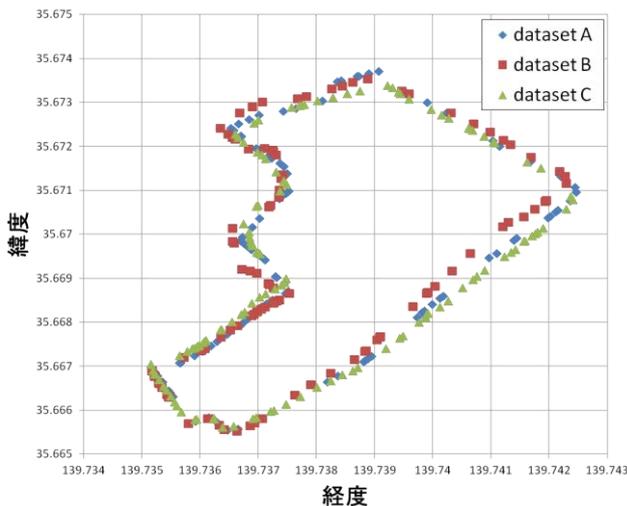


図 4. 各データセットの走行軌跡とチャンク分節点の分布

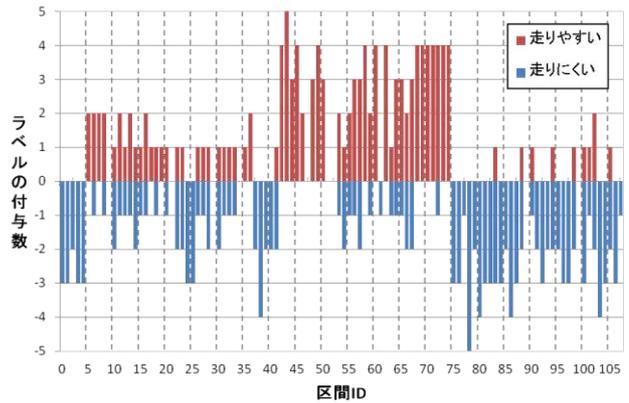


図 5. ラベル付与数の分布(dataset A)

チャンクとしてまとめ上げることで運転時系列データを分節する。

本実験では、表1に記す8種類のラベルを付与した。ラベル付け作業者へ説明した付与基準も合わせて記す。また、使用したデータ諸元を表2に示す。また、各走行データの走行軌跡とチャンクの分節点の分布を図4に示す。

本実験では二重分節解析器によって抽出されたチャンク毎にラベルを付与した。そして、特に「走りやすい」/「走りにくい」の二つのラベルに着目し、区間ごとの分布を評価した。さらに、各ラベルの付与数が過半数を超えた区間を各々「走りやすい」/「走りにくい」区間として抽出し、二つの区間を構成するセグメントの分布の関係性を評価した。

4. 2. 実験結果

4. 2. 1. アノテーション作業負荷の評価

本実験では、ラベル付け作業者には作業内容とアノテーションツールの使い方を15分程度で説明した。初回作業時の作業時間は1周分(10分程度)のデータに対して30~60分程度であった。作業時間の比較のため、1人の作業者にELAN[2]を利用して同等のアノテーションを実施してもらった。当該作業者の作業時間は、提案手法を用いた場合には約45分であったのに対し、ELANを用いた場合には約60分を要した。また、今回と同じ形式のデータに対して、別のアノテーションツールを利用してアノテーションを行った際には、約130時間分のデータに対して約1445時間の作業時間を要していた。これらの事実から、提案手法によってアノテーション作業の負荷を低減できていると言える。

4. 2. 2. 区間ごとのラベル付与数の分布の評価

区間ごとのラベル付与数の分布を図5に示す。図5はdataset Aにおいて、「走りやすい」/「走りにくい」の各ラベルの付与区間と付与数を示す。グラフの横軸は区間ID、縦軸はラベルの付与数(「走りにくい」のラベルはマイナスで表示)を示す。グラフより、区間ID42~50や68~74は「走りやすい」区間、区間ID75~84は「走りにくい」区間、区間ID7~33や53~57などは走りやすさが人によって異なる区間であることがわかる。

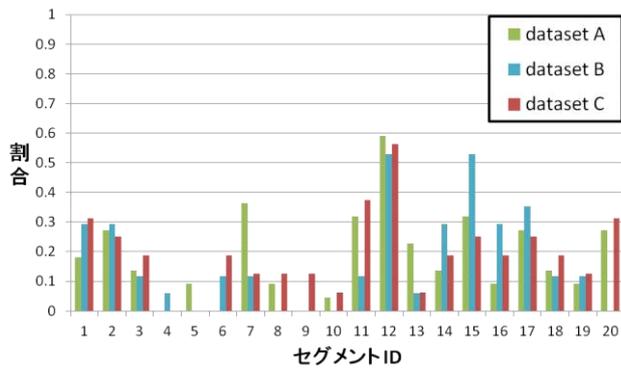


図6. 走りやすい区間のセグメントの分布

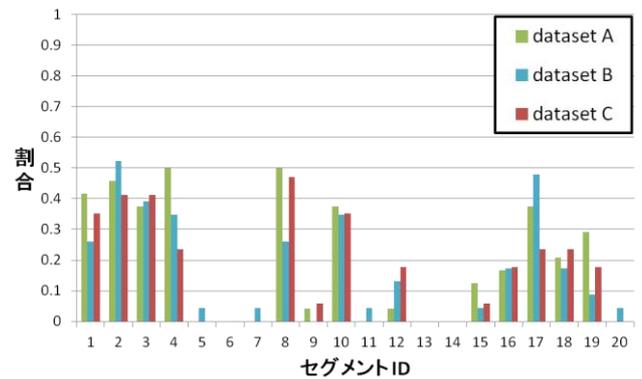


図7. 走りにくい区間のセグメントの分布

表3. ラベル間、dataset間のセグメント分布の比較(ユークリッド距離)

	走りやすい_A	走りやすい_B	走りやすい_C	走りにくい_A	走りにくい_B	走りにくい_C
走りやすい_A	0.00	0.37	0.19	1.46	1.02	1.07
走りやすい_B	0.37	0.00	0.33	1.28	0.95	1.01
走りやすい_C	0.19	0.33	0.00	1.24	0.90	0.86
走りにくい_A	1.46	1.28	1.24	0.00	0.19	0.14
走りにくい_B	1.02	0.95	1.24	0.19	0.00	0.16
走りにくい_C	1.07	1.01	0.90	0.14	0.16	0.00

4. 2. 3. ラベルごとのセグメントの分布の評価

次に、「走りやすい」のラベルが過半数の作業者によって付与されたチャンク区間のうち、各セグメントが含まれていた区間の割合を dataset ごとに集計した。結果を図6に示す。同様に、「走りにくい」のラベルに対して集計した結果を図7に示す。これらの図より、「走りやすい」区間と「走りにくい」区間には各々の区間を良く示すセグメントがあることが分かる。例えば、セグメントID12,15は「走りやすい」区間には頻繁に現れるが、「走りにくい」区間にはほとんど表れないセグメントだとわかる。一方で、セグメントID8,10は「走りにくい」区間には頻繁に現れるが、「走りやすい」区間にはほとんど表れないセグメントだとわかる。また、セグメントの分布はラベルの種類には大きく影響されるが、datasetの違いにはあまり影響されないとと言える。

ラベル間、dataset間のセグメント分布を定量的に比較するために、各 dataset の各ラベル区間のセグメント分布間のユークリッド距離を求めた。その結果を表3に示す。この表より、異なる dataset 間の同じラベル区間のセグメント分布は、異なるラベル区間のセグメント分布に類似していることがわかる。したがって、二重分節解析の結果推定される、ある区間のセグメントの分布から、その区間が「走りやすい」区間なのか「走りにくい」区間なのかを分類することができると考えられる。

5. まとめと今後の課題

本稿では、二重分節解析によって分節した運転時系列データのためのアノテーション手法を提案した。提案手法を用いることで、従来のアノテーション方法では難しかったストリームデータへのアノテーションを軽微な負担で実施できる。また、「走りやすさ」や「危険度」な

どの主観的なアノテーションを始点と終点が明確な形で付与できる。評価実験によって、提案した手法を用いて行うアノテーション結果の有用性を示した。

今後は、本稿で提案したアノテーション手法を含んだ形での運転時系列データの収集・解析体制を構築し、運転支援などのアプリケーションで運転時系列データの解析結果を利用可能な枠組みを構築することが課題となる。

参考文献

- [1] K. Nawa et al., "Cyber Physical System for Vehicle Application," Proceedings of IEEE CYBER 2012.
- [2] <http://tla.mpi.nl/tools/tla-tools/elan/>
- [3] <http://www.anvil-software.org>
- [4] K. Takeda et al., "International Large-Scale Vehicle Corpora for Research on Driver Behavior on the Road," IEEE TITS, Vol.12, Issue 4, pp.1609-1623, 2011.
- [5] 鹿島ら, "クラウドソーシングと機械学習," 人工知能学会誌, Vol. 27, No.4, pp.381-388, 2012.
- [6] N. Kawaguchi et al., "CIAIR In-Car Speech Corpus: Influence of Driving Status," IEICE Transactions on Information and Systems. Vol. E88-D, No. 3, pp.578-582, 2005.
- [7] T. Taniguchi et al., "Semiotic prediction of driving behavior using unsupervised double articulation analyzer," Proceeding of IEEE IV 2012.
- [8] 谷口ら, "運転時系列のベイズ二重分節解析によるチャンク抽出," 計測自動制御学会論文集, Vol. 49, No. 11, 2013.