別添4

平成３０年度　厚生労働科学研究費　補助金
（政策科学総合研究事業（臨床研究等ICT基盤構築･人工知能実装研究事業）
（分担）研究報告書

副作用評価のための機械学習用トレーニングデータの作成および トレーニングデータを用いた試行的副作用評価

研究分担者　潮田　明　国立研究開発法人産業技術総合研究所・人工知能研究センター・招聘研究員

研究要旨

【目的】今年度は副作用評価のための機械学習用トレーニングデータの作成およびトレーニングデータを用いた試行的副作用評価を先行して実施した。

【方法】スティーブンス・ジョンソン症候群（SJS）の副作用症例報告書から、PMDA における評価において副作用として認められた報告書（正例）100 件と、副作用と認めるのに十分な所見あるいは情報がないと判断された報告書（負例）100 件を抽出し、5 分割交差検証を行った。正負判定には、PMDA の専門委員による１次評価の結果を用いた。また複数の専門家からのヒアリング結果および「重篤副作用疾患別対応マニュアル」をもとに９つの素性を機械学習用の素性として選択した。今回は人手による素性の抽出を行い、抽出結果をもとに副作用自動判定の評価を行った。機械学習器には２値分類において汎化能力が高く、精度もトップクラスにある Support Vector Machine (SVM)、および同様に高精度が期待でき更に素性の評価が容易な Maximum Entropy Classifier(MEC)を使用した。

【結果】副作用判定精度に関しては、MEC を学習器として用いた評価において 86.0%の精度を達成した。3 名の専門員の判定精度の概算値が 80%～90%であることから、専門家と遜色のない精度が得られ、手法の有効性が確認できた。

【考察】特徴量抽出以降の工程に関しては、ＡＩの確信度上位５０％の症例に絞れば、正例、負例のそれぞれについてＡＩの精度は９７％～９８％と非常に高く、評価対象の自動絞り込みによる作業効率化への適用可能性が示された。また、「正例（あるいは負例）の可能性が高い」症例の自動分別により、重要案件から優先的に評価することが可能と考えられる。

【今後の計画】2019 年度前半は副作用評価に人工知能を応用するための基盤整備を主として実施する。また、並行して、2018 年度に副作用として認められた報告書（正例）100 件と、副作用と認めるのに十分な所見あるいは情報がないと判断された報告書（負例）100 件を用いて、試行的に構築した人工知能を用いた副作用判定モデルに、さらに 1200 症例のデータを追加してモデルの改良を行い、副作用評価に用いる人工知能の更なる精度向上を目指す。

A．研究目的

副作用を迅速かつ客観的に評価するための人工知能を活用した副作用症例報告評価技術の開発を本研究課題の最終目標として、本分担研究においては、人工知能による汎用性が高い副作用評価を行うための基盤整備として症例報告書からの自動アノテーションを活用した副作用に特徴的な素性（特徴量）の抽出を行う前半部と、得られた素性と素性値から機械学習により副作用判定を行う後半部に分けて研究を行っている。今年度は当初前半部を遂行する予定であったが、個人情報管理の実務的制約が厳しく、副作用症例報告書そのものの内容を扱う前半部よりも個人情報保護への配慮がより軽減される後者の方を先に進めた方がより効率的に研究を進められると判断し、今年度は副作用評価のための機械学習用トレーニングデータの作成およびトレーニングデータを用いた試行的副作用評価を先行して実施した。

B.研究方法

スティーブンス・ジョンソン症候群（SJS）の副作用症例報告書から、PMDA における評価において副作用として認められた報告書（正例）100 件と、副作用と認めるのに十分な所見あるいは情報がないと判断された報告書（負例）100 件を抽出し、5 分割交差検証を行った。正負判定には、PMDA の専門委員による１次評価の結果を用いた。また実際に PMDA において副作用評価を行っている複数の専門家からのヒアリング結果および「重篤副作用疾患別対応マニュアル」をもとに表１に示す素性を機械学習用の素性として選択した。今回は人手による素性の抽出を行い、抽出結果をもとに副作用自動判定の評価を行った。今回の副作用判定は正例と負例を判別する２値分類であるため、２値分類において汎化能力が高く、様々な適用分野において精度もトップクラスにある Support Vector Machine (SVM)、および同様に高精度が期待でき更に素性の評価が容易な Maximum Entropy Classifier(MEC)を機械学習器として使用した。

C．研究結果

得られた副作用自動判定の精度を表 2 に示す。表１の素性と素性値の組み合わせを用いたものを「特徴セット１」、素性番号 1, 2 のそれぞれについて２値の素性値に分解した場合の素性と素性値の組み合わせを「特徴セット２」とした。

Radial basis function(RBF)カーネルと特徴セット２の組み合わせが 83.5%と最も高い精度を示しており、いずれの条件においても８割強の精度が得られた。また、全 200 例を学習データとして用いた MEC で学習されたモデルにおけるそれぞれの素性に対応する重みを比較し、どの素性が副作用判定において重要であるかを見てみると、素性番号 1，9，2，6，4，5，3，7, 8 の順であり、SJS の診断基準でもある粘膜病変や壊死性障害に基づくびらん・水疱といった症状に関する記述の有無が副作用判定において重要であることが示唆された。また、「発熱」は SJS の診断において重要な判断基準とされているが、副作用有りと判定された 100 例中、実際に「発熱」の所見が認められたのは 56 例であったことからも、判断基準としてはやや弱いと考えられた。また、MEC によって学習されたモデルが正解判定およびモデルによる判定双方に与える確率を比較し、モデルからみて信頼度の低い正解判定を調べてみると、今回信頼度が

表１　評価実験で用いた素性のセット

素性番号	素性	素性値	判断基準
1	皮膚粘膜移行部の広範かつ 重度な粘膜病変	2	有り
		1	有り(やや軽度)
		0	なし
2	皮膚の汎発性の紅斑に伴う表皮の 壊死性障害に基づくびらん・水疱	2	有り
		1	有り(やや軽度)
		0	なし
3	発熱	1	有り
		0	なし
4	病理組織所見	1	有り
		0	それ以外
5	皮膚科によるSJS診断	1	有り
		0	なし
6	被疑薬のDSLT検査	1	検査有り(陽性)
		0	検査有り(陰性)
		-	検査なし
7	併用薬のDSLT検査	1	検査有り(陽性)
		0	検査有り(陰性)
		-	検査なし
8	報告書テキスト内に被疑薬投与後の症状発現が記載	1	有り
		0	なし
9	報告書テキスト外テーブルより 被疑薬投与後の症状発現と判断	1	できる
		0	できない

表２　副作用自動判定の精度(%)

正解セット		１人の評価者	
学習器		特徴セット1	特徴セット2
SVM	線形カーネル	80.5	80.5
	RBF(ガウシアン)カーネル	83.0	83.5
Maximum Entropy Classifier		81.5	82.0

表 3　３名の評価者の一致度(Cohen のκ係数)

	A	B	C
A		0.293	0.248
B	0.293		0.287
C	0.248	0.287	

Kappa	一致度
0.00-0.20	低い
0.21-0.40	やや低い
0.41-0.60	中程度
0.61-0.80	かなり高い
0.81-1.00	ほぼ一致

表 4　新旧判定結果の対応表

新正解 / 旧正解	正	負
正	80	20
負	20	80

表 5　200 件全件についての３名の評価者の一致度(Cohen のκ係数)

	A	B	C
A		0.659	0.531
B	0.659		0.582
C	0.531	0.582	

表 6　新しい正解を用いた副作用自動判定の精度(%)

正解セット		1人の評価	3人評価の平均
学習器		特徴セット2	特徴セット2
SVM	線形カーネル	80.5	84.0
	RBF(ガウシアン)カーネル	83.5	84.0
Maximum Entropy Classifier		82.0	86.0
評価者 A			90.0
評価者 B			93.0
評価者 C			86.0
「1人の評価者」			80.0

平均:89.7%
(GSから自分自身の寄与を除くと平均79.3%)

特に低かった事例のなかには、専門家から見てモデルの判断でも間違いとは言えない事例が散見されたことから、モデルのエラー分析が専門家の判断基準の明確化に活用できる可能性も期待された。

評価結果の分析

上記の評価実験では、機械学習モデルの判定と正解判定の間で判断が割れたケースの中には、専門家の間でも判断が割れそうなケースが散見された。またいくつかの症例に関しては、学習済みの AI モデルが判定候補に付与する確率が、モデルが正解判定に付与する確率を大幅に上回っていた。そこで正解判定の検証のために、MEC によって学習された AI モデルが誤った事例 36 件につき、新たに 3 名の PMDA 専門委員（以下 A,B,C とする）により個別に正負判定を行ってその結果を多数決によりまとめた判定を新たな正解としてもとの正解と差し替えて 5 分割交差検証による副作用評価の再実験を行ったところ、前回 AI モデルが誤ったと判定された 36 件中 23 件が正解と判定された。すなわち新たな正解を真の正解と仮定した場合、AI 判定と元の人手判定が異なっていた 36 件中 23 件について AI 判定の方が正しかった事になる。これを以って直ちに AI 判定の方が人手判定より正確であるとは決して言えないが、少なくとも AI が間違える症例については人間の判定にも少なからず揺れが生じるという事が言える。表 3 に再判定を行った 36 件についての 3 名の評価者の一致度(Cohen のκ係数)を示す。いずれの評価者ペアに関してもκ係数は 0.3 未満であり、評価者間の判定の揺れが大きいことが伺える。

正解データの再構築と副作用評価の再実験

上記検証実験により、判定の難しいケースでは評価者間に相当量の判定の揺らぎがあることが明らかになったため、正負データ 200 件全件につき、A,B,C3 名による正負判定の再評価を行った。A,B,C による判定は個別に行い最終判定を機械的に 3 名の判定結果の多数決により行って判定結果を gold standard（GS）とした。表 4 に新旧判定結果の対応表を示す。200 件中 160 件については新旧で判定が変わらず、20 件については正判定が負判定に変わり、残りの 20 件については負判定が正判定に変わった。結果的に新しい正解(GS)の基に正例 100 件および負例 100 件の評価用データが再構築された。表 5 に 3 名の評価者の一致度を示す。前述の 36 件の評価者間一致度に比べて大幅に上昇しており、このことからも前述の 36 件は人間にとっても評価が難しい事例であったことが分かる。副作用判定精度に関しては、MEC を学習器として用いた評価において 86.0%の精度を達成した。表 6 に GS を用いて副作用評価の再実験を行った結果を示す。GS を正解とした時の３名の専門員の判定精度を求めたところそれぞれ 90.0%、93.0%、86.0%となり AI の精度がかなり人間の精度に近づいていることが分かる。また当初の「1 人の評価」の精度を GS を基準に図ると 80.0%となり、AI の精度よりかなり下回っていることが分かった。但しここで「1 人の評価」と呼んでいるのは、1 人の個人がすべて評価したという意味ではない。副作用の１次評価においては１つの症例を１人の専門員が評価するが、同種の副作用の評価をすべて特定の個人が評価するわけではないため、「1 人の評価」は複数専門員が異なる症例を評価した結果の総和になっている。

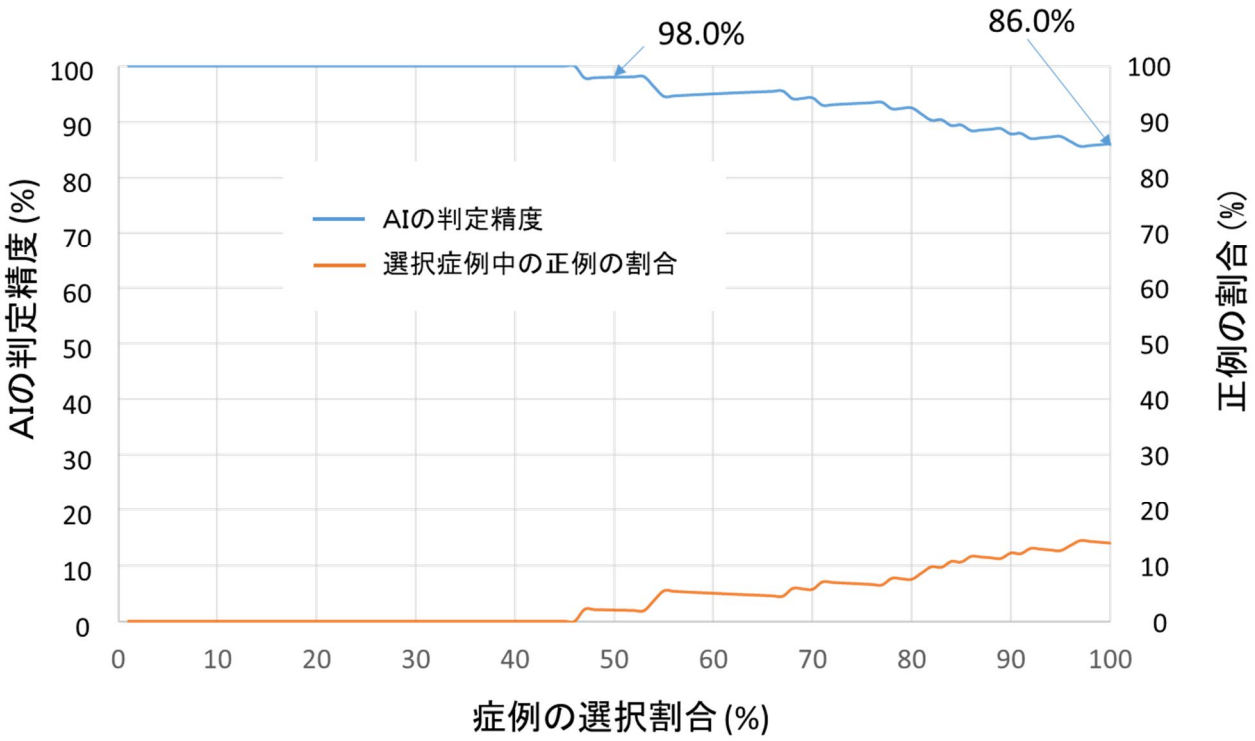

図１　確信度上位の負例判定症例の判定精度

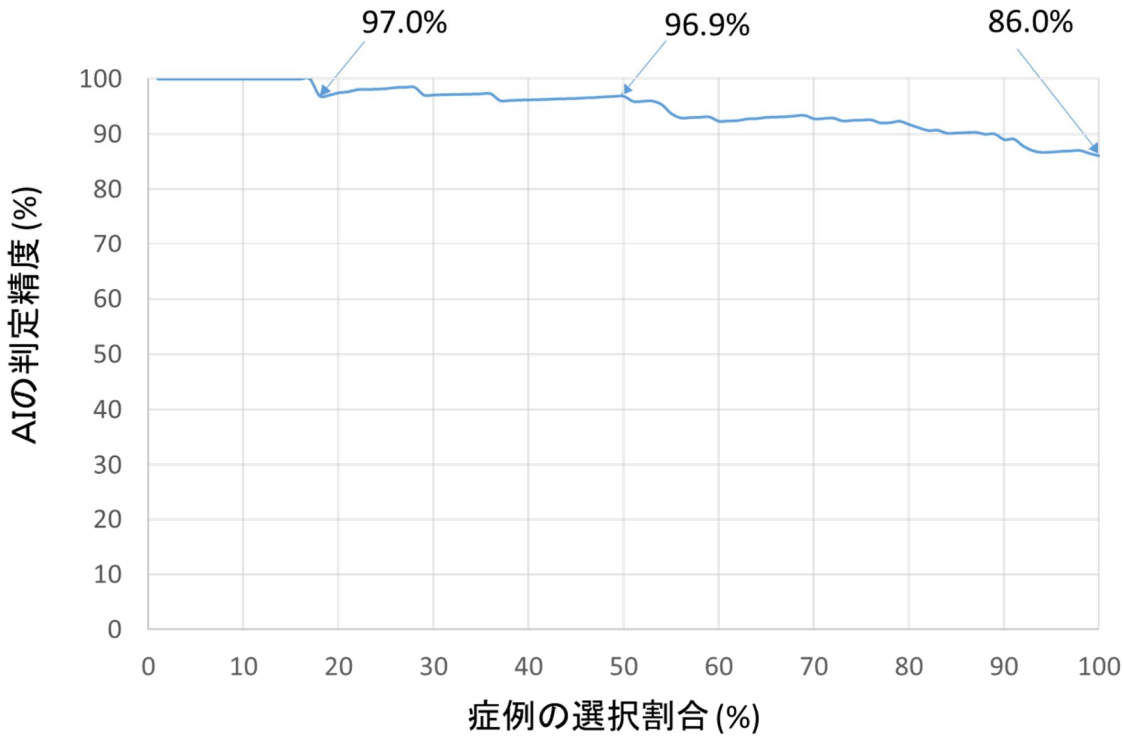

図２　確信度上位の正例判定症例の判定精度

D．考察

特徴量抽出以降の工程に関しては、MEC を学習器として用いた場合 86.0%の精度を達成した。3 名の専門員の判定精度の概算値が 80%～90%であることから、専門家と遜色のない精度が得られ、手法の有効性が確認できたといえる。しかしながら現場で AI を活用しようとした場合、86.0%の精度ではまだ安心して使えるレベルだとは言えない。そこで、AI を現場で役立てて行くための方法の１つとして、AI の確信度、すなわち「ＡＩが自分自身の判定結果に付与する確率」の活用について考察を行った。

図１は AI の確信度と負例判定症例の判定精度の関係を示したものである。横軸は AI が負例と判定した症例（合計 100 件）について確信度の高い順に症例を並べて、特定の確信度以上の症例を選択したときの、選択された症例の数の割合を示し、縦軸は選択された症例に対するＡＩの判定精度を示す。確信度上位４６%の症例に関しては、AI はすべて正解しており、確信度上位５０%の症例の判定精度は９８%であった。図 2 は AI の確信度と正例判定症例の判定精度の関係を示したものである。AI が正例と判定した症例については、確信度上位半分の症例を選択した場合の精度は約97%であった。これらの結果から、AI の確信度の高い結果だけを利用することにより、高精度の判定が実現できる可能性のあることが分かった。また、「正例（あるいは負例）の可能性が高い」症例の自動分別により、重要案件から優先的に評価することが可能であると考えられる。例えば、ＡＩの確信度上位５０%の症例に絞れば、正例、負例のそれぞれについてＡＩの精度は９７%～９８%と非常に高く、評価対象の自動絞り込みによる作業効率化への適用可能性が示された。

E．結論および今後の研究計画

今年度は副作用評価のための機械学習用トレーニングデータの作成およびトレーニングデータを用いた試行的副作用評価を先行して実施した。特徴量抽出以降の工程に関しては専門家の判定精度（８０%～９０%）と遜色のない８６%の精度が得られ、手法の有効性が確認できた。　また AI を現場で役立てて行くための方法の１つとしてAIの確信度の活用が有望であることが分かった。たとえば「正例（あるいは負例）の可能性が高い」症例の自動分別により、重要案件から優先的に評価することが可能と考えられる。また、ＡＩの確信度上位５０%の症例に絞れば、正例、負例のそれぞれについてＡＩの精度は９７%～９８%と非常に高く、評価対象の自動絞り込みによる作業効率化への適用可能性が示された。

2019 年度の研究計画

当初 2019 年度に実施を予定していた「後半部分」に相当する副作用評価のための機械学習用学習データおよび学習モデルの作成を今年度前倒しで開始することができたため、その分来年度の前半は「前半部分」に相当する辞書の構築およびアノテーション用学習データおよびアノテーション用学習モデルの作成に注力することができる。

2019 年度 Q1～Q2 にはまずアノテーション用教師データおよび辞書の作成を行う。2018 年度の小規模サンプルを用いた副作用判定の試行的評価により副作用判定用機械学習に有効な特徴量の洗い出しがある程度行えたため、2019 年度はこれらの特徴量に関連の深い表現（**副作用関連表現**）を中心にアノテーション作業を行う。また、**他の副作用への汎用性を考慮した基盤整備**を目的として、本研究の対象とする副作用に直接関連しない薬剤も含めて症例報告中に出現する**薬剤関連表現**のアノ

テーションも行う。薬剤関連表現の候補として、薬剤名、投与量、投与経路、投与頻度、投与期間、更には投与理由などを検討する。定型的な略語や同義語による表記ゆれを解消するために、薬剤名は、日本標準分類（JSCC）、副作用名は ICH 国際医薬品用語集日本語版（MedDRA/J）、疾患名は国際疾病分類第 10 版（ICD-10）を用いる。また、カルテ文章に頻出する、辞書類などではカバーし切れない非定型な略語や同義語、誤記などによる表記ゆれを解消するために、アノテーション済みテキストから抽出した文脈付き用語と、日本語論文・添付文書並びに Web から収集した大量のテキストとを組み合わせた半教師あり学習により、副作用に関連する多様な表現を抽出し、表記ゆれ解消のための辞書を作成する。作成された辞書とアノテーションデータを対応付けることでデータの標準化を図る。

また Q1～Q2 には並行して副作用評価のための機械学習用教師データの作成を行う。2018 年度の 200 症例分に加え、更に 1200 症例分の副作用評価のための機械学習用教師データを作成する。２値判定の機械学習用データとしては 1400 症例は十分なデータ量と考えられる。

Q2～Q3 にはアノテーション用教師データからアノテーション操作を機械学習できる学習モデルを構築する。また並行して、特徴量を軸にしたエラー分析を通じて逐次的に特徴量と学習モデルの改良を行い機械学習の精度向上を図る。そして Q4 には「前半部」と「後半部」を繋げた副作用評価用機械学習モデルの構築と精度評価を行う。症例報告書テキストから特徴量抽出を行う機械学習と特徴量に基づいて副作用判定を行う機械学習とを結合させて、症例報告書から自動で副作用判定を行う AI モデルを構築する。1400 症例分の副作用評価済みのデータテーブル（素性：素性値ペアのテーブル）を副作用評価用 AI 学習用トレーニングデータとし、残りの約 3000 症例の副作用評価を実施する。3000 症例の副作用症例報告の中から、200 症例程度を抽出し、PMDA の副作用判定の専門家に副作用評価を依頼し、それと AI による副作用判定結果を比較することで、AI による副作用判定の客観的評価を行う。また、その結果をフィードバックし、AI による副作用判定の精度の向上を図る。

F．健康危険情報

該当なし

G．研究成果

1．論文発表

該当なし

2．学会発表

潮田明、今任拓也、森谷純治、斎藤嘉朗、松永雄亮、沼生智晴、見田活、阿川英之、関口遼：人工知能を活用した副作用症例報告書の試行的評価．第 5 回日本医療安全学会学術総会（2019.2）