

一般口演 | 医療データ解析

一般口演11 機械学習

2021年11月20日(土) 09:10 ~ 11:10 G会場 (2号館3階232+233)

[3-G-1-08] ランキング学習アルゴリズムによる診断支援システム - 医師の鑑別診断プロセスと機械学習との融合 -

*宮地 康彦²、鳥越 恵治郎¹、石井 修²（1. 医療法人恵真会 鳥越医院, 2. 臨床推論実用化研究会）

*Yasuhiko Miyachi², Keijiro Torigoe¹, Osamu Ishii²（1. 医療法人恵真会 鳥越医院, 2. 臨床推論実用化研究会）

キーワード：Computer-Aided Decision Support System, Diagnosis Reminder, Clinical Problem-Solving, Machine Learning, Learning to Rank

患者の自他覚症状・理学的所見・臨床検査結果・画像検査結果等から順位付き疾患リストを提示する診断支援システムを開発した。

主な目的は、内科医、総合診療医の臨床推論の支援である。

医師の臨床推論のプロセスは、SOAP(Subjective → Objective → Assessment → Plan)の繰り返しにより、鑑別疾患をリストアップし、最終的に確定疾患を絞り込む。

これまで提案されている診断支援システムの症例データベースは、教科書、症例報告の要旨等の最終的な症状・所見等と確定疾患のみを登録して構築している。また、疾患リストを提示するアルゴリズムは、ベイジアンネットワークでの確率計算、機械学習(分類、回帰)等である。

しかし、従来の手法では、入力する症状・所見等と症例データベースの症状・所見との相関性が高くないと、的確な推論結果は得られない。さらに、従来の手法は、確定疾患のみを対象としており、鑑別疾患のリストは、推論に活用されていない。

筆者らは、これらの問題を解決するために、機械学習の1つであるランキング学習(Learning to Rank; LTR)アルゴリズムによる診断支援システムを開発した。本診断支援システムによる推論結果は、従来の手法と比較して、高い評価結果を得た。

筆者らは、「ランキング学習アルゴリズムによる推論は、SOAP、Clinical Problem-Solving等、医師の鑑別診断プロセスとの親和性が高い」と考えている。

ランキング学習アルゴリズムによる診断支援システム

－ 医師の鑑別診断プロセスと機械学習との融合 －

宮地 康彦^{*2}、鳥越 恵治郎^{*1}、石井 修^{*2}

^{*1} 医療法人恵真会 鳥越医院、^{*2} 臨床推論実用化研究会

Computer-Aided Decision Support System based on LTR algorithm - Collaboration of a clinician and the machine learning in the differential diagnosis -

Yasuhiko Miyachi^{*2}, Keijiro Torigoe^{*1}, Osamu Ishii^{*2}

^{*1} Torigoe-Iin, ^{*2} The Society for Computer-Aided Decision Support System

OBJECTIVE: To support the differential diagnosis for internists and general practitioners, we have developed the "diagnostic reminder system" to infer "ranking list of possible diseases". Diagnostic algorithms of conventional "diagnostic support systems" differ greatly from clinician's differential diagnosis process.

METHODS: The algorithm for machine learning is Learning to Rank(LTR). The evaluation function is "NDCG", and the loss function is "Approximate NDCG". The performance of differential diagnosis was evaluated against Case Records of Massachusetts General Hospital published in the New England Journal of Medicine (NEJM).

RESULTS: Performance of inference: the LTR algorithm has higher than the conventional algorithms. Performance of differential diagnosis: in cases with characteristic symptoms, the LTR algorithm performs almost same as the conventional algorithm; in cases without characteristic symptoms, the LTR algorithm performs better than that.

CONCLUSIONS: The "diagnosis reminder system based on LTR algorithm" has a high affinity with differential diagnosis process of clinician. This system is useful for the following points. 1) Diagnosis reminder of rare diseases. 2) Differential diagnosis for difficult-to-diagnose cases without characteristic symptoms.

Keywords: Computer-Aided Decision Support System, Diagnosis Reminder, Clinical Problem-Solving, Machine Learning, Learning to Rank

1. 緒論

1.1 概要

"患者の自覚症状、他覚症状、理学的所見、臨床検査結果、画像検査結果等"(以下、症状等)から、"可能性のある疾患の順位付きリスト"を提示するシステムを開発した。

目的は、内科医、総合診療医の臨床推論(鑑別診断)の支援である。

1.2 背景

鑑別診断を構築して、検査を選択し、その結果を解釈することは、すべての医師にとって重要な技術である。

医師は、以下のステップで鑑別診断をおこなう。¹⁾

Step 1: 問題を同定する。

Step 2: 鑑別診断の枠組みをつくる。

Step 3: 鑑別診断を整理する。

Step 4: 鑑別診断を絞り込む。

Step 5: 病歴と診察所見を用い、考えられる診断を探索する。

Step 6: 鑑別診断に優先順位をつける。

Step 7: 仮説を検証する。

Step 8: 新たなデータに基づき、鑑別診断に再び優先順位をつける。

Step 9: 新たな仮説を検証する。

医師は、上記のステップを繰り返して、新たなデータ(症状等)を得て、鑑別診断(鑑別疾患)の絞り込みをおこない、最終的な確定疾患を決定する。

本論の診断支援システムは、"Step 4"、"Step 6"と"Step 8"

について、医師の鑑別診断を支援するものである。

診断支援システムの最重要ポイントは、症例データベースと推論アルゴリズムである。

症例データベースは、通常、"医学書、医学論文、症例報告等"(以下、文献等)から、疾患に関する情報をテキストマイニングして構築される。優れた文献等や、特に、Clinical Problem-Solving(CPS)形式の文献等は、上記の鑑別診断プロセスについて論述し、確定疾患だけでなく鑑別疾患に関する情報も論述している。これらの情報を効果的にテキストマイニングできれば、経験のある臨床医の鑑別診断プロセスを反映した優れた症例データベースとなる。

また、推論アルゴリズムも、上記の症例データベースを効果的に使用し、かつ、経験のある臨床医の鑑別診断プロセスを可能な限りシミュレートできれば、網羅性、順位等の推論性能、そして、推論結果の解釈可能性と説明可能性等の点から、優れた推論アルゴリズムとなる。

医師の鑑別診断プロセスの視点から、既存の診断支援システム^{2), 3), 4)}を評価した場合、以下の"問題点"がある。

1) 最終的な確定疾患と(複数の)鑑別疾患との関連性に関する情報を有効利用していない。

2) 診断支援システムは、"可能性のある疾患の順位付きリスト"の出力が最終的な目的である。

いずれの診断支援システムも、個々の疾患ごとの機械学習によりスコアを計算し、その結果をスコア順にソートして出力している。

しかし、推論結果の網羅性、順位等の評価は、機械学習の評価関数ではなく、経験のある臨床医による主観的な評価にとどまり、客観的かつ自動化されていない。

また、その評価結果は、的確に推論モデルにフィードバックされていない。

つまり、既存の診断システムは、症例データベースの構築と利用、推論アルゴリズムの点から、医師の鑑別診断プロセスを十分に反映していない。

2. 開発目的

医師の鑑別診断プロセスは、"鑑別診断の絞り込みと優先順位付けの繰り返し"である。

"背景"の"問題点"を解決し、医師の鑑別診断プロセスとの親和性の高い診断支援システムとして、ランキング学習アルゴリズムによる支援システムを開発した。

ランキング学習(Learning to rank, LTR)は、機械学習の手法であり、推論結果の網羅性、順位の最適化(訓練)、評価を目的としている。

これを利用した診断支援システムは、鑑別診断の絞り込み(= 網羅性)、優先順位付け(= 順位)を最適化(訓練)し、評価関数によって客観的に評価する。

3. システム概要

3.1 データベースの構築

症状等マスタ、疾患マスタ、症例データ等のデータベースは、筆者らが、文献等から、疾患に関する情報をテキストマイニングして構築した。

3.1.1 症状等マスタ

症状等は、自覚症状、他覚症状の他に、理学的所見、臨床検査結果、画像検査結果等を含む。

臨床検査結果は、例えば、血糖値は、検査結果値(例: 空腹時血糖値 200mg/dL)ではなく、高血糖として表現する。同様に発熱、脈拍、呼吸数、白血球などの定量的数値は採用せず、高低、多寡についてのみ表現する。

画像検査結果は、例えば、胸部 X 線検査は、画像データ(例: DICOM 形式)ではなく、所見(例: 胸部レ線異常、骨レ線異常)などと表現し、異常の質・量は特殊な所見(腫瘍、無気肺、胸水、気胸の存在など)を除いて採用しない。

症状等マスタのコード数は、583 である。

コード体系、名称は、独自仕様である。

入力時の検索を支援するために、1 コードに、複数の症状等の定義や意味などが記載されている。

3.1.2 疾患マスタ

疾患マスタのコード数は、1,000 である。

コード体系、名称は、独自仕様である。

鑑別診断を支援するために、1 レコードに、鑑別疾患として近縁の疾患が並記されており、疾患の総数は 6,460 である。

3.1.3 症例データ

CPS 形式などの鑑別疾患プロセスが論述されている文献等からは、確定疾患に関する情報だけでなく、鑑別疾患に関する情報もテキストマイニングする。

鑑別疾患のスコアは、鑑別診断の進行(= 症状等の数の増加)に応じて、設定する。

なお、最終的な確定疾患となる疾患は、常に、任意のスコア α (例: 10)を重み付けして(= 足して)、設定する。

症例データのデータ数(= 症例数)は、23,395 である。

(表 1)

症状データは、データフレーム形式変換する。

(表 2)

3.2 システムの開発環境、実行環境

機械学習ライブラリは、TensorFlow、TensorFlow Ranking を使用した。

TensorFlow Ranking は、TensorFlow 用の、ランキング学習用のライブラリである。⁵⁾

評価用の実行環境は、クラウド コンピューティングサービス(Google Cloud Platform 等)を利用した。(表 3)

表 3 システムの開発環境、実行環境

種類	環境
プログラム言語	Python 3.8.x
機械学習ライブラリ	TensorFlow 2.4.x TensorFlow Ranking 0.3.x
ウェブ アプリケーション フレームワーク	Flask 2.x.0
クラウド コンピューティング サービス	Google Cloud Platform App Engine AI Platform Firebase Authentication

3.3 ニューラルネットワークの構成

評価関数は、DCG(Discounted Cumulative Gain)を正規化した NDCG(Normalized DCG)を採用した。

DCG、NDCG は、ランキング学習の推論結果(網羅性、順位)を直接評価する。⁶⁾

損失関数は、Approximate NDCG(ApproxNDCGLoss)を採用した。

Approximate NDCG は、NDCG を微分可能に変換した関数である。⁷⁾(表 4)

表 4 ニューラルネットワークの構成

項目	設定
入力層	ユニット数 583 (症状等コード数)
中間層	層数 1 ユニット数 1024
出力層	ユニット数 1000 (疾患コード数)
損失関数	ApproxNDCGLoss
評価関数	ndcg ndcg@1, ndcg@3, ndcg@5 mean_squared_error(mse)

4. システム評価と考察

システム評価のため、他の診断支援システムとの比較評価をおこなった。

比較対象は、筆者らが先に開発・評価した"I-DR"である。⁴⁾

I-DR は、本システムのベースとなったシステムであり、機械学習のアルゴリズムは、回帰であり、損失関数は、MeanSquaredError(MSE)である。

損失関数以外の訓練データ、評価データ、ニューラルネットワークの構成等は、両システムとも同一条件とした。

4.1 学習曲線、評価関数による評価と考察

4.1.1 評価方法

症例データをランダムに分割して、訓練データ(80%)、評価データ(20%)を生成し、訓練、評価をおこなった。

評価関数(ndcg)の値は、1に近いほど、推論性能が高いと評価できる。

4.1.2 評価結果

学習(訓練)が収束する学習回数(epochs)は、ランキング学習のほうが少ない。ただし、ランキング学習は、演算量が多いため、実際の学習時間は、回帰のほうが短い。

(図 1)

ランキング学習では、学習回数を増やすと、訓練データの ndcg の値は 1.0 に収束するが、評価データの ndcg の値は 0.7 前後でほぼ一定となる。

最終的な ndcg の値は、訓練データ、評価データともに、ランキング学習のほうが、常に高い。(表 5)

4.1.3 考察

ランキング学習は、回帰と比較して、常に推論性能が高い。

しかし、ランキング学習は、過学習の傾向があり、汎化性能はあまり高くない。

本システムの開発の過程で、様々なパラメータのチューニング等を実施したが、汎化性能はほとんど向上しなかった。

よって、現時点において、推論性能を向上させる最も確実な手法は、症例データの蓄積であり、様々なパターンの症例を登録することにより、全体的な推論性能の向上の期待ができる。

4.2 実際の症例による評価と考察

4.2.1 評価方法

症例は、NEJM の最新の症例報告から引用した。

それぞれの症例について、最終推論結果と、鑑別診断の進行(= 症状等の追加)による確定疾患の順位の変化を評価した。

診断支援システム上の疾患は、代表的な疾患のみを表示する。

なお、公正を保つため、これらの症例は、評価に使用した症例データベースには、登録していない。

4.2.2 評価結果と考察

4.2.2.1 症例 1: 急性間欠性ポルフィリン症⁸⁾

最終的な確定疾患の推論順位は、ランキング学習が 2 位で、回帰が 1 位である。

確定疾患の推論性能としては、大きな差は無い。(表 6)

ランキング学習の推論結果は、除外すべき疾患¹⁰⁾である虫垂炎、鉛中毒がリストアップされており、医師の鑑別診断に有用な情報を提示している。

鑑別診断の進行による確定疾患の順位の変化は、両アルゴリズムともに、本疾患の特徴的な症状等¹⁰⁾である低ナトリウム血症と肝機能異常を追加した段階で、上位にリストアップされ、その後の変化は少ない。(表 7)

よって、特徴的な症状等を有する症例に関しては、両アルゴリズムで推論性能に大きな差はないと考える。

4.2.2.2 症例 2: 急性 HIV-1 感染症⁹⁾

HIV 感染は初感染時に急性髄膜炎症状を呈することがあり、本症例もこれに該当する。よって、急性ウイルス性髄膜炎も、確定疾患に準ずる疾患として評価する。

最終的な確定疾患の推論順位は、ランキング学習が、1 位: 急性 HIV 感染症、2 位: 急性ウイルス性髄膜炎であり、回帰では両疾患とも 20 位以下である。

確定疾患の推論性能は、ランキング学習のほうが高い。

(表 8)

鑑別診断の進行による確定疾患の順位の変化は、ランキング学習では、早い段階から急性ウイルス性髄膜炎が常に上位にリストアップされている。(表 9)

本症例の症状等は、いずれも、単独の症状等としては、他の多くの疾患でもありうる症状等である。

よって、特徴的な症状等が少ない症例に関しては、ランキング学習のほうが、推論性能が優れている可能性が高いと考える。

5. 結論

内科医、総合診療医の臨床推論(鑑別診断)の支援を目的として、"患者の自覚症状、他覚症状、理学的所見、臨床検査結果、画像検査結果等"から、"可能性のある疾患の順位付きリスト"を提示するシステムを開発した。

本システムは、経験のある臨床医の鑑別診断プロセスとの整合性が高い機械学習のアルゴリズム(ランキング学習)を採用している。

本システムは、機械学習の面からの性能評価と、臨床医学の面からの実症例による評価のいずれにおいても、既存の診断支援システムと比較して、高い推論性能を示した。

筆者らのこれまで知見から、本システムには、以下の将来性、可能性があると考えている。

- 1) 頻度が少ない疾患の思い出し
- 2) 特徴的な症状等が少ない診断困難な疾患の鑑別診断
- 3) 推論結果の解釈可能性と説明可能性の明確化

筆者ら、引き続き、理想の診断支援システムを追求していく。

6. 参考文献

- スコット・スターン, アダム・シーファー, ダイアン・オールトカーム, 竹本 毅. 考える技術 第4版. 日経BP, 2020
- 永井良三. 第38回医療情報学連合大会(第19回日本医療情報学会学術大会), [4-A-2-2] 人工知能による総合診療診断支援システムの開発
- 鳥越恵治郎. コンピュータ診断支援ツール -病名思い出しから病名絞り込みへ-. 日本医事新報 2006; 4310: 73-75
- 栗山裕, 曾田祐民, 矢野愛華 ら. 内科におけるコンピュータ利用診断支援システムによる診断改善. 岡山医学会雑誌 2019; 29-34
- Rama Kumar Pasumarthi, Sebastian Bruch, Michael Bendersky, Xuanhui Wang. Neural Learning to Rank using TensorFlow. ICTIR 2019.
- Discounted cumulative gain.
https://en.wikipedia.org/wiki/Discounted_cumulative_gain (cited 2021-08-01)
- Hamed Valizadegan, Rong Jin, Ruofei Zhang, Jianchang Mao. Learning to Rank by Optimizing NDCG Measure.
- Thomas W. Fredrick, M.D., Manuel B. Braga Neto, M.D., Ph.D., Daniel O. Johnsrud, M.D., Michael Camilleri, M.D., and Victor G. Chedid, M.D. Turning Purple with Pain. N Engl J Med. 2021; 385: 549-554
- Robert H. Goldstein, M.D., Ph.D., William A. Mehan, M.D., Bailey Hutchison, M.D., and Gregory K. Robbins, M.D. Case 24-2021: A 63-Year-Old Woman with Fever, Sore Throat, and Confusion. N Engl J Med 2021; 385: 641-648
- 難病情報センター: ポルフィリン症(指定難病 254).
<https://www.nanbyou.or.jp/entry/5546>
(cited 2021-08-01)

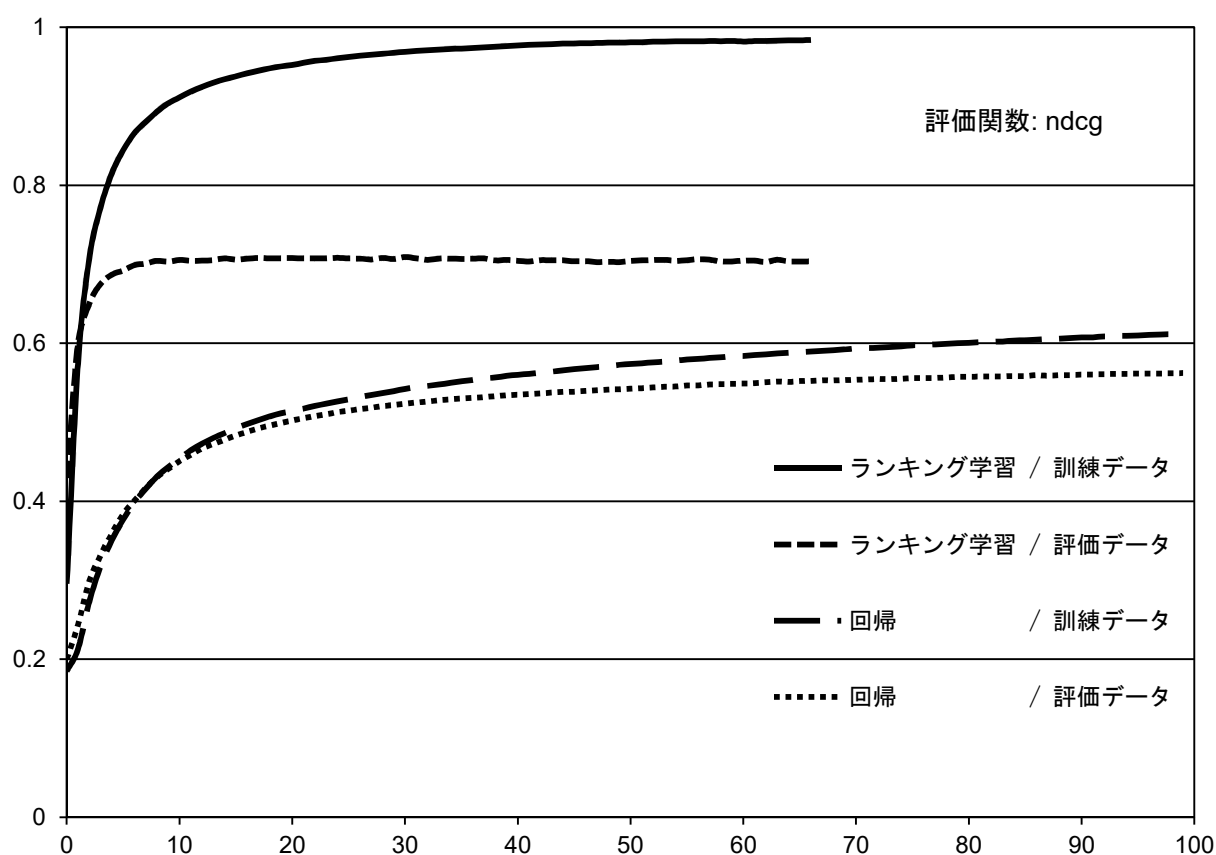


図1 学習曲線

表5 評価関数による評価

アルゴリズム	損失関数	評価関数(訓練データ)				
		ndcg	ndcg@1	ndcg@3	ndcg@5	mse
ランキング学習	ApproxNDCGLoss	0.9855	0.9910	0.9872	0.9850	34.3957
回帰	MSE	0.6148	0.3653	0.4495	0.4859	0.1522

アルゴリズム	損失関数	評価関数(評価データ)				
		ndcg	ndcg@1	ndcg@3	ndcg@5	mse
ランキング学習	ApproxNDCGLoss	0.7039	0.5094	0.5849	0.6138	34.2786
回帰	MSE	0.5624	0.3031	0.3031	0.4165	0.2280

表 1 症例データの例

No.	code	症状等	No.	code	score	疾患
2	s31	腫脹	2	d31	5	うっ滞性皮膚炎
2	s34	発赤	2	d32	30 + α	蜂窩織炎
2	s35	圧痛	2	d33	5	深部静脈血栓症 (DVT)
2	s32	腫脹: 急性	2	d34	10	Baker 嚢胞破裂
2	s36	超音波検査: 血管: -	2	d35	10	リンパ路閉塞を生じる腫瘍
2	s39	抗菌剤治療: +				

*参考文献 1)をもとに作成

表 2 症例データフレームの例

No.	s1	...	s31	s32	s33	s34	s35	s36	s37	s38	s39	...	s51	d1	...	d31	d32	d33	d34	d35	...	d51
2	0		1	1	0	1	1	1	0	0	1		0	0		5	30+ α	5	10	10		0
	←												→	←								→

*参考文献 1)をもとに作成

表 6 症例 1: 急性間欠性ポルフィリン症 / 最終推論結果

入力した症状等	ランキング学習	回帰
	順位 予測疾患	順位 予測疾患
痛み(腹痛)	1 門脈閉塞症・門脈血栓症・門脈炎	1 急性間欠性ポルフィリン症
便秘	2 急性間欠性ポルフィリン症	2 コレラ・病原性大腸菌感染症
低ナトリウム血症	3 糖尿病性昏睡切迫状態	3 内臓破裂
腹部圧痛	4 低ナトリウム血症	4 線維筋痛症(結合織炎)
肝機能異常	5 熱中症	5 急性(ウイルス性)胃腸炎(食中毒)
頻脈	6 鉛中毒症(殆んど慢性)	6 摂食障害
不安感	7 内臓破裂/出血(血腫)/梗塞	7 薬物/大麻/嗜好品の乱用による症状群
口渇感	8 急性(化膿性)虫垂(憩室)炎	8 激しい脱水症
腹鳴(蠕動増多)	9 アセトアミノフェン中毒症	9 過敏性腸症候群・機能性消化管障害
脱水症	10 覚醒剤中毒症	10 十二指腸または胃穿孔・大腸穿孔等
錯乱状態		

表 7 症例 1: 急性間欠性ポルフィリン症 / 鑑別診断の進行による確定疾患の順位の変化

追加した症状等	ランキング学習	回帰
	順位: 急性間欠性ポルフィリン症	順位: 急性間欠性ポルフィリン症
痛み(腹痛)	→ 20 以下	20 以下
便秘	→ 13	20 以下
低ナトリウム血症	→ 2	2
腹部圧痛	→ 2	20 以下
肝機能異常	→ 3	2
頻脈	→ 1	1
不安感	→ 1	1
口渇感	→ 2	1
腹鳴(蠕動増多)	→ 1	1
脱水症	→ 2	1
錯乱状態	→ 2	1

表 8 症例 2: 急性 HIV-1 感染症 / 最終推論結果

入力した症状等	ランキング学習		回帰	
	順位	予測疾患	順位	予測疾患
発熱	1	急性 HIV-1 感染症	1	急性 Q 熱感染症/慢性 Q 熱感染症
頭痛	2	急性ウイルス性髄膜炎	2	リンパ球性脈絡髄膜炎
咽頭痛	3	西ナイル熱/脳炎	3	口唇ヘルペス
意識不鮮明	4	一酸化炭素中毒症	4	咽後膿瘍・咽頭後部感染症
悪寒戦りつ	5	インターフェロンの副作用	5	急性咽頭炎/急性咽頭喉頭炎
筋肉痛	6	猫ひっかき病	6	A 型(流行性)肝炎,
嚥下痛	7	(特発性)正常圧水頭症	7	多発性神経障害
咽頭喉頭異常感	8	多発性神経障害	8	慢性疲労症候群/筋痛性脳脊髄炎
失語	9	A 型(流行性)肝炎	9	感染性心内膜炎・慢性心内膜炎
失行	10	脳腫瘍/(異型)髄膜腫/大脳膠腫瘍	10	狂犬病(恐水病)
倦怠感				
全身の脱力感				
食欲不振				
体重減少				
認知症				

表 9 症例 2: 急性 HIV-1 感染症 / 鑑別診断の進行による確定疾患の順位の変化

追加した症状等		ランキング学習		回帰	
		順位: 急性 HIV-1 感染症	順位: 急性ウイルス性髄膜炎	順位: 急性 HIV-1 感染症	順位: 急性ウイルス性髄膜炎
発熱	→	20 以下	20 以下	20 以下	20 以下
頭痛	→	20 以下	4	20 以下	20 以下
咽頭痛	→	9	4	20 以下	20 以下
意識不鮮明	→	20 以下	1	20 以下	8
悪寒戦りつ	→	20 以下	1	20 以下	8
筋肉痛	→	20 以下	2	20 以下	20 以下
嚥下痛	→	4	1	20 以下	20 以下
咽頭喉頭異常感	→	20 以下	4	20 以下	20 以下
失語	→	20 以下	5	20 以下	20 以下
失行	→	20 以下	5	20 以下	20 以下
倦怠感	→	20 以下	5	20 以下	20 以下
全身の脱力感	→	20 以下	4	20 以下	20 以下
食欲不振	→	14	2	20 以下	20 以下
体重減少	→	1	3	20 以下	20 以下
認知症	→	1	3	20 以下	20 以下

以上