

一般口演 | 病院情報システム

一般口演8

病院情報システム

2019年11月22日(金) 14:50 ～ 16:50 F会場 (国際会議場 3階中会議室302)

[2-F-2-04] 疾患別機械学習モデルの管理を実現したシステム基盤の検討

○廣田 健一¹、小山内 尚² (1. 札幌医科大学, 2. 株式会社富士通北陸システムズ)

キーワード : Machine learning, Machine learning models management, Cloud technology, Container technology

札幌医科大学附属病院（以降、当附属病院）では、機械学習を活用した糖尿病治療における経口血糖降下薬の処方最適化に関する研究を進めているが、その他疾患に関する学習モデル作成の研究も順次検討している。このように、診療情報データの利活用に機械学習を適用することがさらに増加することが考えられるが、大学病院における全診療情報データをパブリッククラウドに展開することは、個人情報保護、レスポンス速度確保などについてさらなる検討を要する。よって、当附属病院を含めた医療機関において機械学習基盤を安価に構築可能であり、かつ将来性を考慮したシステム方式を考える必要がある。これらの要件を満たすためには、複数の疾患に関するデータセット、学習モデルを同時に実行可能なこと、他病院においても実現可能になるように汎用性が高いこと、構築費用が安価であること、今後のパブリッククラウドへの展開を見据えた構成であること、などが重要である。実現するにあたり、パッケージソフトではなく、複数のオープンソースソフトウェア（以降、OSS）を組み合わせることを選択し、代表的な OSS としては、コンテナ管理基盤に Kubernetes、データ分析基盤に Jupyter を導入した。Jupyter 上に scikit-learn、TensorFlow 等の現時点でデファクトスタンダードと考えられる機械学習フレームワークが動作する仕組みの導入を行った。その際、外部ネットワーク回線に繋がらない環境にて OSS を利用するため、独自のパッケージレポジトリを構築し、各種 OSS の最新化が実現可能な仕組みを実現した。なお、本システムは、コンテナレベルにおいて互換性があるため、将来的には複数の学習モデルをパブリッククラウドへスムーズな移行を可能としており、さらに医療情報システムとのデータ連携も見据えたインターフェイスについても検討を進めている。

疾患別機械学習モデルの管理を実現したシステム基盤の検討

廣田 健一^{*1}、小山内 尚^{*2}

^{*1} 札幌医科大学、^{*2} 株式会社富士通北陸システムズ

The effort for system infrastructure that realizes disease-specific machine learning models management

Kenichi Hirota ^{*1}, Hisashi Osanai ^{*2}

^{*1} Sapporo Medical University, ^{*2} Fujitsu Hokuriku Systems Limited

Recently, there are an increasing number of cases where machine learning is applied to analyzing electronic health record in medical institutions. In the cases, it is necessary to effectively manage and operate a plurality of disease-specific machine learning models, however the approach has not yet been clearly presented. This research aims to present a machine learning platform specialized for medical institutions and a platform that looks ahead to collaboration with medical information systems. For this purpose, considering the requirements in the clinical field, we examined the unified management and operation of the disease-specific data sets, machine learning models, outputs and their relationship. As a result, the machine learning platform is realized with a combination of multiple open source software, such as using Kubernetes as the container management and Jupyter as a data analysis tool and so on. Finally, realized diabetes treatment and research on other diseases.

Keywords: Machine learning, Machine learning models management, Cloud technology, Container technology

1. 背景と目的

札幌医科大学附属病院(以下、当附属病院)では、ルールベースによる人工知能(以下、AI)の推論利用と言える B 型肝炎再活性化防止システム^{*1}、機械学習を活用した糖尿病治療における経口血糖降下薬の処方最適化に関する AI による学習モデル構築の研究^{*2}に加え、他疾患に関しても同様の研究を順次進めているところである。このように、医療情報システムに蓄積された診療情報データの利活用を進めることは医療機関では必須となっており、分析結果からの知見を活用することで、さらに医療の質や医療安全の向上が期待されている。近年では、そのデータ分析手法の一つとして機械学習を選択することが増加しており、当附属病院に限らず、全国的な潮流になっていることは周知の事実である。当附属病院においては、複数の機械学習による学習モデルの作成研究を進めているが、医療分野の機械学習における技術方法として、必要となるデータセット、学習モデル、アウトプットの効果的な一元管理、運用方法が示されていないのも事実であり、これらを実現するシステム基盤の構築は、医療機関における機械学習利用の重点ポイントと考えている。

現在の機械学習の技術レベルを活用して構築される学習モデルは、全ての疾患に適用できるという汎用的学習モデルである「汎用型 AI」ではなく、疾患別学習モデルである「特化型 AI」と言われている機械学習の技術である。そのため、これら疾患別学習モデルの結合体が一元管理のもと運用され、患者の疾患状況に応じて、動的にこれら学習モデルからのアウトプットが医療者へ提供されることが臨床現場における機械学習利用の重点ポイントであり、それらの詳細方式については未だに世界的に論じられてはいない。さらに、これら疾患別の学習モデルからアウトプットされる結果を医療情報システムに適用することが重要であり、臨床現場から求められている。なぜなら、臨床現場においては、効率的かつ効果的な医療を行うために、医療情報システムの活用は必須であり、そこに

機械学習の結果を用いた何らかの“サジェスション情報”を利用可能であることが、機械学習の臨床貢献と言えるであろう。しかし、この特化型 AI の統合という技術を医療分野において完成させているという事例はなく、やはり特化型 AI の単独利用が中心になっている。本研究においては、複数の特化型 AI を利用した学習モデルなどを、複数のユーザが利用可能なシステム基盤の実現について論じている。

また、他業種における機械学習の利用形態は、Google Cloud Platform 等の大手クラウドベンダのパブリッククラウド上で行われていることを主流としている。しかしながら、大学病院における診療情報データをパブリッククラウドに展開するためには、個人情報保護やデータ流出などのセキュリティリスクに加え、診療に耐えうるレスポンス確保などの課題があり、これらの解決には時間を要する。よって、当附属病院を含めた医療機関において機械学習基盤を構築するためには、以下の要件を満たすことが必要と考える。

- (1) 複数の疾患に関するデータセット、学習モデル、アウトプットが一元管理され、運用可能なこと
- (2) 臨床現場において実際に実用可能であること
- (3) 今後のパブリッククラウドへの展開を見据えたシステム構成であること

前述した医療機関における診療情報データの利活用への機械学習の適用においては、セキュアであり、かつ複数の特化型 AI を有効に管理、運用することが必要であり、それらについては未だに明確に提示されてないため、本研究では、これら医療機関に特化した機械学習基盤についての提示を目的とする。また、このシステム基盤は医療情報システムとの連携を見据えた基盤を考えており、さらに将来的において、クラウド環境へのスムーズな展開が可能であることも目的としている。

2. 方法

ここでは、医療機関における機械学習基盤に求められる各要件に関する要素技術について述べる。

2.1 疾患別のデータセット、学習モデル、アウトプットの一元管理、運用の実現

機械学習が適用される診療情報データには、検査データ、処方データ等の数値データ、放射線等の画像データが存在しており、当附属病院における、それら研究におけるデータ管理方法について記述する。機械学習を活用した糖尿病治療における経口血糖降下薬の処方最適化に関する AI による学習モデル構築の研究では、糖尿病治療において重要である指標である HbA1c が 7.0%未満を維持できている期間をコントロール期間、HbA1c が継続的に 7.0%以上となっている期間をアンコントロール期間と定義している。また、経口血糖降下薬の処方により、HbA1c の改善が見られた患者を「正例データ」、改善の見られなかった患者を「負例データ」と呼称し、「正例データ」は、アンコントロール期間が存在し、かつ、コントロール期間が 365 日以上である患者と定義した。また、「負例データ」はコントロール期間が存在しない、あるいは、コントロール期間が 365 日に満たない場合と定義した。下記に、正例データおよび負例データのイメージ図を示す。

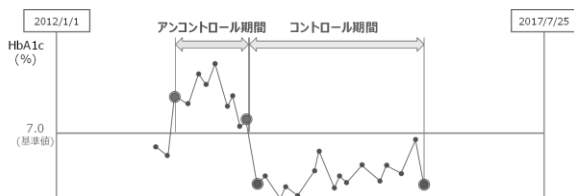


図 1 正例データのイメージ



図 2 負例データのイメージ

このような定義に従い機械学習の適用に向けたデータセットを作成しているが、これらのデータ抽出・加工には、データウェアハウスやビジネスインテリジェンス(以下、DWH、BI)を使用している。これらから、必要となる検査データ、処方データ等の数値データを DWH や BI 等から抽出し、成形化されたデータセットとして機械学習基盤に提供している。また、胸部 X 線画像から特定の疾患をスクリーニングする学習モデルの構築を目的とした研究においては、医用画像管理システム(以降、PACS)にてデータ管理されている画像データから、匿名化、成形化された画像データを作成し、データセットとして、機械学習基盤に提供している。

これらの各研究に必要なデータセット、このデータセットから構築される学習モデル、そこから導かれるアウトプット情報を一元的に管理する必要性は高い。

この機械学習基盤に関するイメージを以下に示す。各データセットにおいては、利用ユーザが異なること、また、データセットである数値データ、画像データが学習モデルの精度と紐づいた形でデータ版数管理機能が求められる。本研究では、検査データ等の数値データに代表される構造化されたデ

ータ、放射線等の画像データに代表される構造化されていない(非構造化データ)を一元管理するために、データ構造を問わないオブジェクトストレージを採用しており、各疾患のデータセットを一元管理し、それぞれの学習モデルとの紐付けを行った上で管理している。また、学習モデル作成の簡易化のため、データセットをオブジェクトストレージに格納後、自動にてハイパーパラメータのチューニングを含む学習処理を動作させている。これら実現の基礎技術面をカバーしているのが Kubernetes^{*3}、Docker^{*4} 等のオープンソースソフトウェア(以降、OSS)を必要に応じて組み合わせることで、コンテナ技術を利用して、コンテナ単位での可搬性により、これらを実現している。

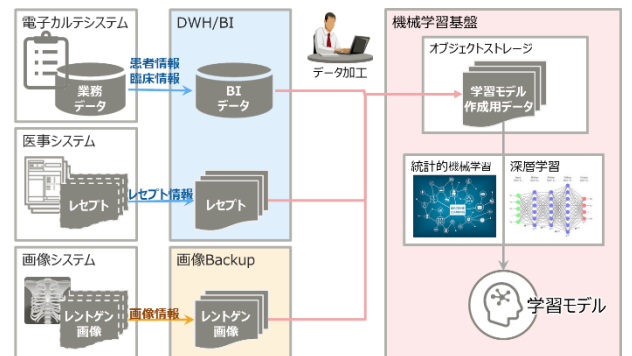


図 3 オブジェクトストレージによるデータの一元管理

2.2 パブリッククラウド展開を見据えた構成

これら機械学習基盤技術は、パブリッククラウド環境と同様の技術を利用可能と考えている。しかし、アプリケーション・プログラミング・インターフェース(以下、API)については、その実現範囲に大きな違いがある。これは、パブリッククラウドにおける API は、クラウドベンダーが保有するデータとそれらのデータから作成された学習モデルにより構成されるため、本研究で述べるシステム基盤では、パブリッククラウドのレベルには、遠く及ばないことは事実である。また、TensorFlow^{*5}、Keras^{*6}等の機械学習フレームワークは、データと組み合わせることでその価値が出るものであることから、その普及のため OSS として広く提供されており、パブリッククラウドと同じレベルのものが使用可能となっている。

これらより、今回の検討では、データ分析ツールに Jupyter^{*7}を導入した。Jupyter 上では scikit-learn^{*8}、Monasca-Analytics^{*9}、TensorFlow 等の現時点でデファクトスタンダードと考えられる機械学習フレームワークが動作する仕組みの導入を行った。

目的別API (SaaS)	オブジェクト認識	テキスト分析	言語変換	感情分析
機械学習向けIaaS/PaaS	TensorFlow	Keras	mxnet	PyTorch
	learn	Caffe	Caffe2	CNTK
	Monasca/Analytics			
コンピュータ・インフラストラクチャ	CPU	GPU		

図 4 パブリッククラウドにおける機械学習基盤構成

3. 結果

検査データ等の数値データ、放射線等の画像データから構成される各疾患のデータセットとそれに紐づけた学習モデルとその精度情報をオブジェクトストレージで、一元管理した。具体例として、機械学習を活用した糖尿病治療における経口血糖降下薬の処方最適化に関する AI による学習モデ

ル構築の研究では、最適な学習モデルとしての評価を AUC 値に加え、正解率と正例の再現率を用いた。AUC 値とは、ROC 曲線下の面積を指し、分類問題における性能の良さを表す指標である。これらの指標値は、現時点で AUC 値 0.925、正解率 94.4%、正例の再現率 73.6%となっているが、この評価に関する精度情報を疾患データセット、そこから作成された学習モデルと紐づけての管理を自動で行っている。また、機械学習基盤のコンテナ管理に Kubernetes、データ分析ツールとして Jupyter 等を利用する等、複数の OSS を組み合わせた構成とすることで、今後のパブリッククラウド展開を捉え、かつ、安価なシステム構築が可能となった。その際、医療機関ならではの課題である外部ネットワークに接続できない環境にて OSS を利用するため、独自のパッケージレポジトリを構築し、各種 OSS の最新化ができる仕組みを実現した。

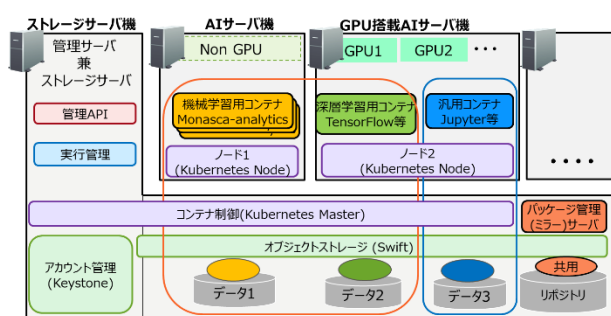


図 5 機械学習基盤の構成

4. 考察

臨床現場において実際に実用可能な機械学習基盤に求められる要件(疾患毎に関連する情報の一元管理、臨床の現場での実用可能性、将来性のある構成)に対して、技術要素とそれらを必要とする背景を整理し、その妥当性を実証した。今後、電子カルテシステム等の医療情報システムと連携した診療情報データ利活用が一層進むことを見据え、一度作成した学習モデルの更新、電子カルテシステム等との連携について検討を行っている。

参考文献

- 1) 廣田健一, 大西浩文, 射場浩介, 千葉弘文. 医療情報システムによる先進的な B 型肝炎再活性化防止対策に関する取り組み. 医療の質・安全学会誌 2018; vol.13 no.2: 130-134.
- 2) 札幌医科大学. 糖尿病治療における経口血糖降下薬の処方最適化に関する AI による学習モデル構築の共同研究を開始. <http://web.sapmed.ac.jp/jp/news/press/jmjbbn000000b6g3.html> (cited 2019-Feb-12).
- 3) Kubernetes community. What is Kubernetes. <https://kubernetes.io/docs/concepts/overview/what-is-kubernetes/>.
- 4) Docker community. Why Docker?. <https://www.docker.com/why-docker/>.
- 5) TensorFlow community. Why TensorFlow. <https://www.tensorflow.org/about/>.
- 6) Keras community. Why use Keras?. <https://keras.io/why-use-keras/>.
- 7) Jupyter community. The Jupyter Notebook. <https://jupyter.org/>.
- 8) scikit-learn community. scikit-learn. <https://scikit-learn.org/stable/#>.
- 9) Monasca-Analytics community. MoNanas – Monasca Analytics Framework overview. <https://github.com/openstack/monasca-analytics#overview>.