

機械学習応用システムの開発に関わって

西村 駿人^{1,a)}

概要：ソフトウェアエンジニアとして、機械学習を利用したシステムの開発チームに参加した経験をまとめる。また開発の取り組みの中で直面した問題を示し、改善につながる提言を示す。

1. はじめに

私は企業の研究所に所属する「ソフトウェア工学の研究者」であり、役割として「ソフトウェアエンジニア」でもあるという認識を持っている。この夏からソフトウェア開発時の問題解決を支援するための機械学習を利用したシステムの研究開発のチームに参加している。

この稿ではまず一つの事例として、私に関わる機械学習応用システム^{*1}の開発について紹介する。次に開発事例の中で、直面した問題を示し、解決のための提言を示す。

2. 携わった機械学習応用システム開発の概略

機械学習応用システムの概略について述べる。プロジェクトで開発する機械学習応用システムは、主に社内での利用を想定しており、システムエンジニアがソフトウェア開発上の問題を解決するための新規のシステムである。具体的な機械学習の問題としては、多クラス分類問題として設定した。

次に開発体制について述べる。プロジェクトのメンバーは、ソフトウェア工学領域の問題を研究開発の対象とする研究所のグループの他に、その問題の対応にあたるシステムエンジニアのグループと、実際にその分野に知見のある社内の有識者のグループという、それぞれの専門家たちが集まるプロジェクトであった。私を含む多くのメンバーは、機械学習応用システムの開発は初めての経験であった。

最後に開発の計画と実施状況について述べる。最初に、我々の抱えるソフトウェア開発時の課題について、その課題を解くべきか、またどのような方法で解くべきかの検討が行われ、機械学習手法を用いたアプリケーションでの解決策が良いのではないかという結論に至った。次に分類問題に利用するためのデータ収集が行われ、メンバーによる

クラス情報のアノテーション作業に充てられた。^{*2}教師データを利用した簡単な分類器を構築し、構築した分類器をテストデータに対して検証することで、性能の確認を行った。その後はプロジェクトに所属しない一部のユーザに対して利用してもらうことでシステム出力の妥当性の確認などを行い、その後社内のシステムエンジニアに利用してもらう予定である。

執筆時は頑健な分類のために、別の学習手法の検討や前処理手法の変更などの作業中で、ウィンターワークショップ開催時は、パイロット運用目前の状態を予定している。

3. 開発中に直面した問題と解決のための提言

以下に3点の問題と、その解決のための提言を示す。

3.1 開発計画を立てることが難しい

多くのメンバーは、初めて機械学習応用システムの開発を行うため、どのように進めればいいのか手探りの状態だったように思う。^{*3}新規のシステムであるため、社外でも利用される直接利益を生み出すシステムと比較すると、事前のビジネス上の KPI (Key Performance Indicator) の設定が困難であった。また、必要なリソースを事前に見積もり、管理するタイプのシステム開発を本業としているであろうシステムエンジニア側からすると、見積もりが難しく試行錯誤をするタイプの開発スタイルはやきもきする進め方であったのではないと思う。

提言としては、既存のシステム開発との進め方に対する違いを認識し、そのベストプラクティスを考えておく必要がある。現在、機械学習応用システムを開発することは、広く一般的にはなっていないと考えており、多くの人がその開発を体験していないと考えている。プロジェクトのメンバーは必要に応じて、機械学習応用システムの開発に関

¹ 株式会社富士通研究所

^{a)} n.hayato@jp.fujitsu.com

^{*1} 本稿での用語は [1] を参照した。

^{*2} 私はここから参加した。

^{*3} 機械学習応用システムの開発経験のあるメンバーによって助けられた。

する書籍を読むことや、簡単な機械学習プログラムを実行してみることで、知見を得ることができる。またその情報をチーム内で共有することが、有用であると感じた。

開発の進め方や体制は、所属団体やプロジェクトによって異なることもあるため、その差分については当事者で差分を発見し、その差分が問題となりうるならば、解決策を発見する必要もあるだろう。そのため過去に機械学習応用システムの開発・運用経験のある人が多ければ多いほうが、当然ながら機械学習応用システムの適用についての成功する可能性は高まると考えている。^{*4}機械学習応用システムの開発プロセスを標準化することは、必要かどうかはわからないが、チームや同じドメインの事業ごとなど、適切な粒度では検討の余地があると考えている。

3.2 教師データの不足に関する問題

プロジェクトが直面した大きな問題としては、利用可能な教師データの不足であった。当初、プロジェクトが使えるデータの多くは、大量のラベルの付与されていないもので、教師データとしてそのまま利用できるかたちのデータは限られている状況であった。さらにラベルのないデータに対して、いかに効率よくラベルをつけるかということに頭を悩ませていた。

提言としては、前節と同様にこれまでのシステム開発との進め方の違いについて知見を持ち備えておくことに加えて、メンバーのデータについての理解があることが重要であると考え、今回の開発では、チームメンバー全員に高品位なデータの作成の方法や、データ作成にかかるコストなどを含めた検討内容などを伝え、さらに当初検討していた手法以外のデータの集め方を検討した。また現状のデータで技術的に実現可能か調べるために、学習器を作成しながら検証を進めていくというやり方をとり、ドメイン知識のある専門家も参加し結果を確認することで作業を進めた。

また稼働中のシステムのログデータの活用し、教師データとする場合もあるだろうが、当初の計画では機械学習しないにしても、適切なログ設計などをしておくことがこれからはさらに大切になってくると考える。機械学習応用システムではない開発においても、機械に利用させる可能性があるという観点を持って開発を行うと、後に利用しやすいのではないかと思う。

3.3 他者と共同作業しやすい仕組みの構築が難しい

最後の問題として、データ解析業務について途中から参加のしやすい状態を作ることの難しさや、他の開発者メンバーと共同しての学習器などの開発が難しいことを挙げる。残念ながら私のプロジェクトでは具体的には分析したリソースが引き継ぎやすい状態になっておらず、なぜその

分析を実施したかなどの試行錯誤の履歴がよい粒度で残っていない。そのため私が参加した当初は過去の内容を調べる必要があり、今では、開発リポジトリを他のメンバーに引き継ぐことは気乗りしない状態である。また他者と協業するために、インタフェースとなるデータのフォーマットも扱いにくい設計であることや、その他にもソースコード以外のリソースの適度なバージョン管理も難しく感じている。

提言としては、システムのユースケース、機械学習の仮説と検証、および、それに付随する目的や条件、ソースコード、出力ログ、データ分析手法、分析結果などをいかに管理するかの方法論を整備し、その環境の整備を実施する必要がある。開発者などのある立場の人にとっては使いやすいツールであっても、ビジネス上の戦略を考える人など立場の異なる人にとっては使いやすいものとは限らない。ステークホルダー全員が利用しやすい環境は何かということについて、ソフトウェア工学の研究者の知見を利用し、整備する必要があると考えている。

4. おわりに

私の経験した開発との中で直面した問題、解決策のアイデアについて述べた^{*5}。現在では手法自体の性能向上や業界自体の盛り上がりもあり、多くのシステムに機械学習手法を組み込んで活用しようという空気感を感じるが、人工知能は万能であるという誤解を持つ人も依然として多いと思われる。「これくらいのリソースがあれば、これくらいのことができる」という一般化は難しいだろうが、多くの人はニュースや新聞記事、技術系の Web サイト等で取り上げられる情報から、どうしても期待感を持ってしまう。

本稿の執筆中に同僚との会話で、「ウォーターフォールモデル」という用語は社内でも伝わるが、機械学習アプリの開発のプロセスを表す用語はあるだろうか、という話になった。機械学習応用システム開発のプロセスについては書籍等はある [2] が、まだまだ広く一般に浸透していないのではと思う^{*6}。もしそうだとしたら、私たちに何ができるだろうか。

以上の経験を基に議論に参加し、今後の活動につなげたい。

参考文献

- [1] 丸山宏：機械学習工学に向けて、日本ソフトウェア科学会第34回大会（2017年度）講演論文集（2017）。
- [2] 有賀康顕，中山心太，西林孝：仕事ではじめる機械学習，オライリー・ジャパン（2017）。

^{*5} 取り組む上での問題点は、人工知能の研究開発を本業とする当社他部署の同僚、また社外の頼れる友人知人たちに相談に乗ってもらい非常に感謝している。

^{*6} あるいは実は既に浸透しており、私だけが取り残されている可能性もあるが。

^{*4} これは特に機械学習応用システムに限った話ではないだろうが。