楽天におけるビッグデータを対象とした機械学習・深層学習の活用事例

平手　勇宇

楽天では，楽天市場をはじめとするさまざまなサービスを提供しており，それらのサービスから日々膨大なデータが生成され続けている．本稿では，楽天が保有するビッグデータを対象とした，機械学習および深層学習の活用事例について紹介を行う．

キーワード：E-commerce，機械学習，Deep Learning，自然言語処理，顧客モデル，レコメンダーシステム

1. はじめに

インターネットサービスあるいは，IoT デバイスの普及により，画像データ，ユーザ行動データ，センサーデータをはじめとするさまざまなビッグデータの取得が比較的容易になってきた．また，GPU や分散ファイルシステムの普及により，大量のデータを蓄積し，それを高速に処理する基盤を容易に構築することができるようになってきた．この二つの要因によって，現在，大量のデータを対象にした機械学習・深層学習の適用が幅広く行われている．

楽天においても，2018 年現在，70 を超えるサービスを運用しており [1]，それらのサービスから商品データやユーザ行動データなど，日々大量のデータが生成されている．それらのデータを基にして，サービス・UX の改善につながるような機能の構築，あるいはマーケティングにつながるような知識を抽出することは非常に重要であることから，自社のデータを対象にした，機械学習・深層学習を適用させる取り組みを数多く推進している．

本稿では，楽天が保有するデータを対象とした，機械学習アルゴリズム，深層学習アルゴリズムの適用事例について紹介を行う．

2. 機械学習・深層学習を利用したプロジェクト例

楽天には，大きく分けて，商品データや商品レビューデータなどの商品に関連するデータ，商品購買履歴データや商品閲覧履歴データなどのユーザ行動に関連するデータの 2 種類が存在する．われわれは，これらのデータを対象とし，多種多様な用途に機械学習・深層学習アルゴリズムを適用させている．

具体的には，図 1 に示すように，(1) 商品を検索しやすくするための商品データの構造化，(2) 多数登録されている商品レビュー・店舗レビューをユーザに見やすく提示することを目的としたレビュー解析，(3) 外部からのアタックを遮断し安全なサービスを提供するためのユーザログインの異常検知，(4) やらせレビューなど不適切な情報を抽出し削除する不正検知，(5) ユーザの購買・閲覧履歴データを基にした商品推薦システム・パーソナライズドクーポンアルゴリズムの構築，(6) 在庫数の最適化のための商品需要予測，(7) マーケティングのためのカスタマーモデルの構築，(8) 多腕バンディットアルゴリズムを用いた AB テストの最適化といったプロジェクトを実施している．

本稿ではその中でも，E-commerce にとって特に重要である商品データの構造化，潜在顧客抽出，レコメンダーシステム・クーポン配信について紹介を行う．

3. 商品データの構造化

楽天市場には，2018 年 12 月現在，約 2 億 6300 万種類の商品が登録されている[1]．その商品データは，楽天市場に出店いただいている店舗によって登録・編集されているため，構造化されていないプレインテキストで書かれた商品データや，誤った商品カテゴリに登録されているなどノイズとなる商品データが数多く存在している．このような商品データは，ユーザによる

ひらて　ゆう
楽天株式会社 楽天技術研究所
yu.hirate@rakuten.com

[1] https://www.rakuten.co.jp

Copyright © by ORSJ. Unauthorized reproduction of this article is prohibited.

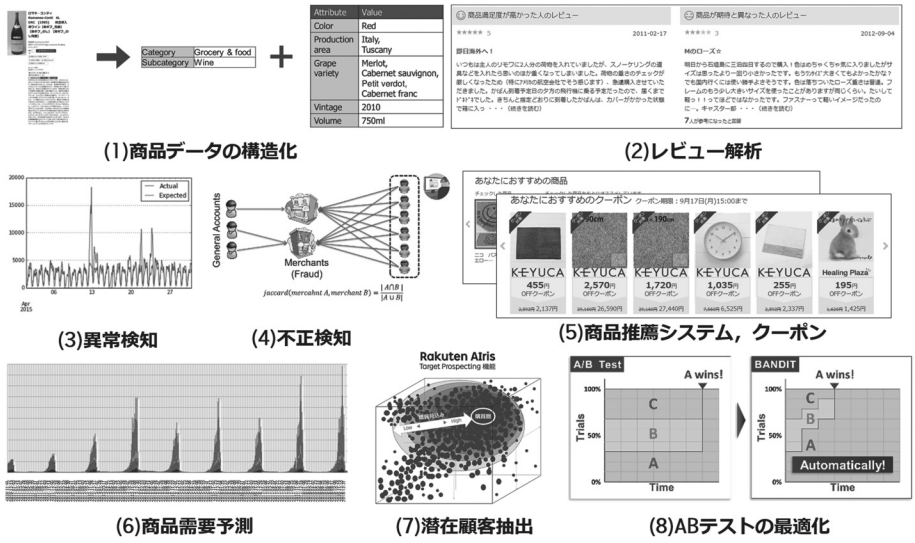

図 1　機械学習・深層学習を利用したプロジェクト例

図 2　商品データの構造化

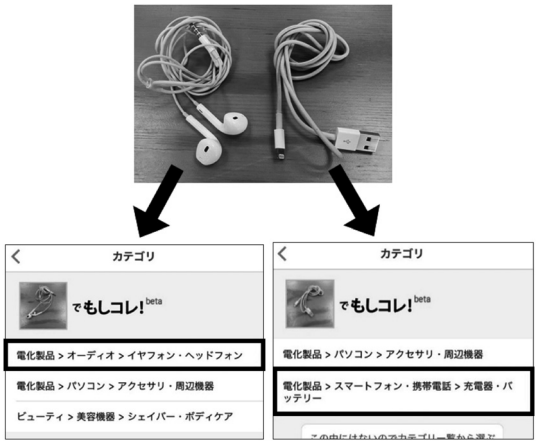

図 3　画像による商品カテゴリ推定

商品検索の利便性を下げる要因となるため，図 2 に示すように商品カテゴリを推定し誤った商品カテゴリに登録されている商品を検知したり [2]，商品の属性・属性値を抽出する [3] ことで，商品の検索性を向上させることが必須となる．

商品カテゴリの推定は，商品タイトル，商品説明文，あるいは商品画像から，当該商品が属する楽天市場の商品カテゴリを推定するタスクであり，機械学習の多クラス分類に相当する．楽天市場の商品データには，すでに商品カテゴリ情報が市場出店店舗によって付与されているため，これらを教師データとしてみなし[2]，分類器を構築する．これまで，商品タイトル・説明文などのテキスト情報を対象とし，Random Forest [4] や，Xgboost [5] のような弱学習器を組み合わせた分類器を適用をさせてきたが，近年では，CNN などの Deep Learning の手法を適用して高い精度の分類器を構築することに成功している [6].

また，図 3 に示すように，CNN を活用して，商品画像からの商品カテゴリ推定についても取り組んでおり，C2C アプリなど，ユーザからアップロードされた商品画像を基に商品カテゴリを推定し，ユーザによる商品カテゴリ指定を補助する機能などに活用をされている．

[2] すべての商品データが正しい商品カテゴリに属しているわけではないので，教師データの中にも一定の確率で誤りが存在する．

 Copyright © by ORSJ. Unauthorized reproduction of this article is prohibited.

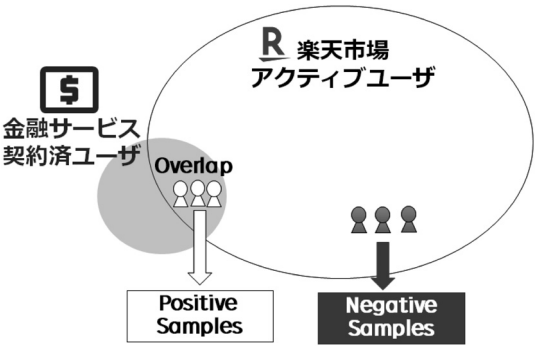

図 4　楽天市場のデータを用いた潜在顧客モデル構築の概念図

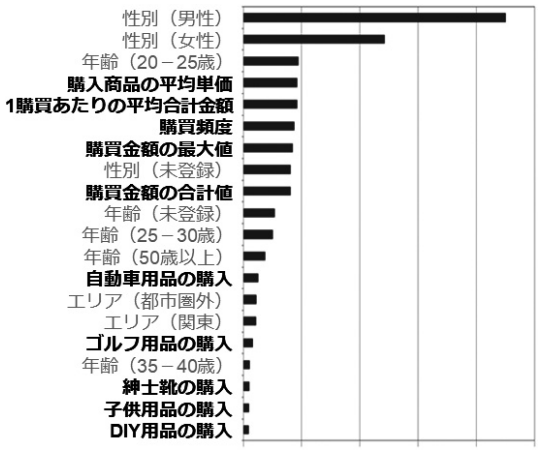

図 5　金融商品向け潜在顧客モデルの重要なユーザ Feature

4.　潜在顧客抽出

楽天市場の商品閲覧履歴データ，商品購買履歴データは，ユーザの趣味・興味を推定するうえで，非常に有効なデータである．また，楽天市場には，約 2000 万人のアクティブユーザ[3]が存在するため，上記の閲覧履歴データ，購買履歴データは多くのユーザを網羅している．この特性を利用することで，楽天市場のデータを利用し顧客モデルを構築することで，任意の商品・サービスに興味をもつであろう潜在顧客を高い精度で抽出することができる．

潜在顧客抽出の具体例を図 4 に示す．図 4 では，楽天グループの金融商品（住宅ローン）の潜在顧客を楽天市場のアクティブユーザの中から抽出する例である．あらかじめ，市場アクティブユーザそれぞれに対して，ユーザ属性情報，楽天市場での購買頻度，商品カテゴリごとの購買の有無などの Feature を生成しておく．そのうえで，図 4 に示すように，当該金融商品をすでに契約しているユーザ集合と，楽天市場のアクティブユーザ集合の積集合を抽出する[4]．この積集合のユーザを Positive サンプル，残りの市場アクティブユーザ集合からランダムにピックアップしたユーザを Negative サンプルとみなし，金融商品の潜在顧客モデルを構築する．

図 5 は，Random Forest [4] を用いて当該金融商品の潜在顧客モデルを構築した際の重要なユーザ Feature である．図 5 では，購買商品の平均単価，1 購買当たりの購買金額の平均，購買頻度など，ロイヤリティが高いユーザであることを示す Feature が重要であると同時に，自動車用品，ゴルフ用品，紳士靴，子供用品，DIY 用品を購入したことのあるユーザであることを示す Feature が重要であるということが自動的に導き出されており，妥当なモデルであるということが示唆される．実際に，この潜在顧客モデルを利用して，メール上での広告を掲載した結果，ランダムにピックアップしたユーザと比較して，広告のクリック率がプラス 49.23%と高い数値を示せており，ターゲティングが有効であったことが確認されている．

上述のスキームは，購入済み・契約済みのユーザと類似したプロパティをもつユーザを潜在顧客とみなし抽出するという仕組みであるため，図 6 に示すように，任意の商品・サービスに適用することが可能である．そこで，下記に示すような汎化・高機能化を施すことによって，広告商品として楽天グループ外にも提供している [7]．

- 楽天会員情報，楽天市場の購買履歴以外のデータも対象として Feature を生成
- RNN や，CNN などの Deep Learning Algorithm をモデル構築に適用

5.　レコメンダーシステム・クーポン配信

商品推薦，あるいはクーポン配信は，基本的な E-commerce のデータ活用方法である．楽天でも，図 7 に示すように，購買履歴データ，閲覧履歴データ，商品データを基にして，商品推薦・クーポン配信の機能を実装している．

Traditional には，相関ルール抽出 [8]，Matrix Factorization [9] などの単一アルゴリズムを適用し，商品推薦，クーポン配信システムを実装をしてきたが，近年では，図 8 に示すように，複数のアルゴリズムの解

[3] 過去 3 カ月以内に一度でもログインをしたことのあるユーザを指す．
[4] 楽天グループでは単一のユーザ ID を付与しているため，複数サービスのユーザ集合を簡単に結合することができる．

Copyright © by ORSJ. Unauthorized reproduction of this article is prohibited.

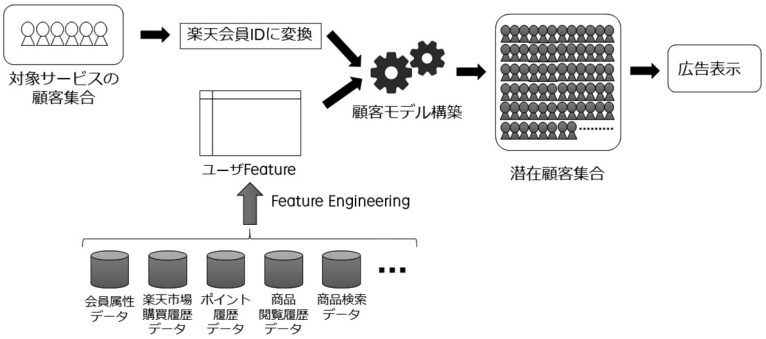

図 6　潜在顧客抽出ソリューションのフレームワーク

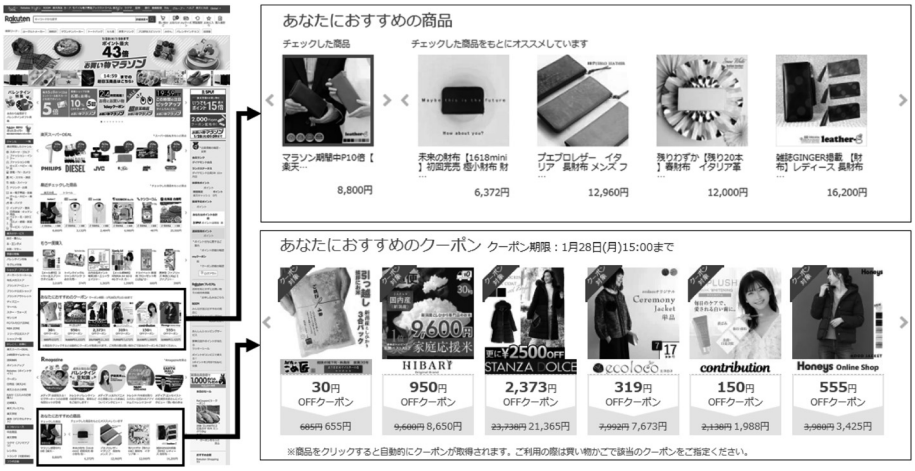

図 7　楽天市場における商品推薦システム，クーポン配信システム

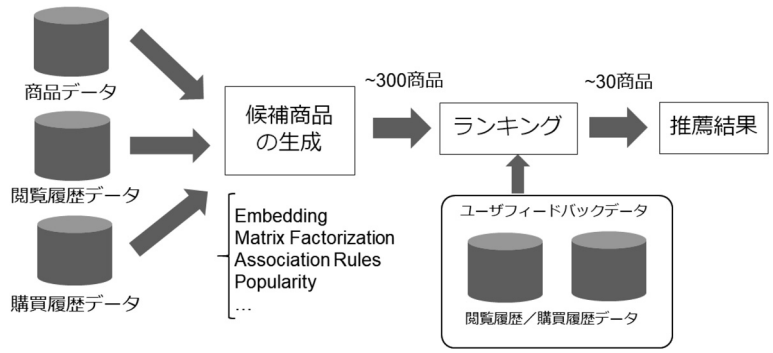

図 8　商品推薦アルゴリズムの構成

を組み合わせて最終的な推薦案を生成する手法を取り入れている．具体的には，商品データ，購買履歴データ，閲覧履歴データを対象とし，古典的な相関ルール抽出，Matrix Factorization のアルゴリズムを適用させたり，商品やユーザなどを固定長の Vector に変換して距離を計算する Embedding [10, 11] を用いた手法 [12]，あるいは RNN や Autoencoder などの Deep Learning アルゴリズムベースの手法 [13, 14] を適用させている．また，単純な商品の人気度による商品推薦も重要なアルゴリズムの一つである．これらのアルゴリズムにより数百の候補を抽出する．そして，直近の購買履歴データ，閲覧履歴データから生成されるユーザフィードバックデータを基に先ほどの数百の候補商品をランキングする．最後に，上位数十商品をピック

 Copyright © by ORSJ. Unauthorized reproduction of this article is prohibited.

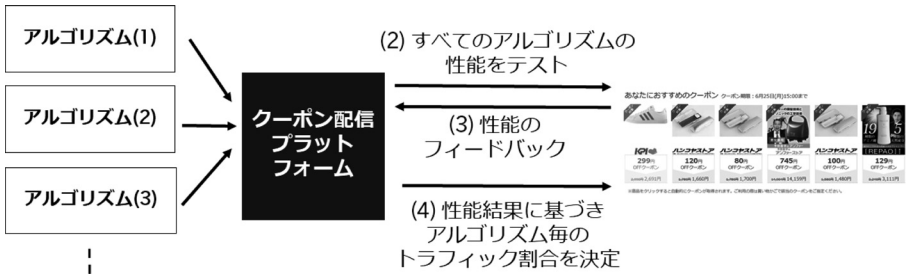

図 9　クーポン配信システムによる最適化

アップすることで最終的なアウトプットとする．

また，アルゴリズムの組み合わせ，ランキング方法，パラメータのセッティングなどをカスタマイズすることによって，極めて多数のレコメンダー手法を定義することが可能である．したがって，単一のレコメンダー手法を優先するのではなく，常に複数の手法を同時に活用しながら，最適な手法を選択する仕組みづくりが重要である．楽天では，図 9 に示すようなクーポン配信プラットフォームを構築し，クーポン配信の最適化も行っている．具体的には，下記に示すようなプロセスを日次で実施している．

1. 複数のクーポン配信手法の設計者が日次で手法を更新し，配信プラットフォームに登録をする．
2. 比較的ユーザのトラフィックが少ない時間帯[5]を選び，各手法を適用させる．
3. CTR（= Click Through Rate），CVR（= Conversion Rate）などを計算することで，各手法のパフォーマンスを評価する．
4. 残りの時間帯で，最も高いパフォーマンスを示した手法にユーザトラフィックの大部分を付与するといったように，評価結果に基づき各手法に与えるユーザトラフィック割合を調整する．

このようにプラットフォームの運用，各クーポン配信手法の性能を客観的に評価することができ，継続的な手法改善を行えるだけではなく，新しいアイデアに基づくクーポン配信手法を容易にテストできるため，システム全体の最適化にも大きく寄与している．

6. おわりに

本稿では，楽天におけるビッグデータを対象とした機械学習・深層学習の活用事例として，商品データの構造化，潜在顧客抽出，およびレコメンダーシステム・クーポン配信の紹介を行った．このほかにも図 1 に示すようなさまざまな事例に機械学習・深層学習を適用させている．

E-commerce が普及するにつれ，物流がボトルネックとなりつつある．オンラインのユーザ行動に関するデータを対象とするだけではなく，倉庫内のオペレーションの最適化，配送センター配置の最適化，配送ルートの最適化など，データの側面から物流の最適化を試みるトピックが多く存在する．今後は，このようなトピックに取り組んでいく予定である．

[5] 2018 年 12 月現在は夕方の時間帯を割り当てている．

参考文献

[1] 楽天株式会社コーポレートレポート 2017，https://corp.rakuten.co.jp/investors/assets/doc/documents/ar_2017_all_jp.pdf（2019 年 1 月 25 日閲覧）
[2] A. Cevahir and K. Murakami, “Large-scale multi-class and hierarchical product categorization for an E-commerce giant,” In *Proceedings of the 26th International Conference on Computational Linguistics*, pp. 525–535, 2016.
[3] K. Shinzato and S. Sekine, “Unsupervised extraction of attributes and their values from product description,” In *Proceedings of the 6th International Joint Conference on Natural Language*, pp. 1339–1347, 2013.
[4] L. Breiman, “Random forests,” *Machine Learning*, **45**, pp. 5–32, 2001.
[5] T. Chen and C. Guestrin, “Xgboost: A scalable tree boosting system,” In *Proceedings of the 22nd Intarnational Conference on Knowledge Discovery and Data Mining*, pp. 785–794, 2016.
[6] Y. Xia, A. Levine, P. Das, G. Di Fabbrizio, K. Shinzato and A. Datta, “Large-scale categorization of Japanese product titles using neural attention models,” In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 663–668, 2017.
[7] Rakuten AIris, https://adsales.rakuten.co.jp/download/media/guide/2019/rakuten_airis-target_prospecting.pdf（2019 年 1 月 25 日閲覧）
[8] R. Agrawal and R. Srikant, “Fast algorithms for mining association rules in large databases,” In *Proceedings of the 20th Intarnarional Conference on Very Large Data Bases*, pp. 487–499, 1994.
[9] Y. Koren, R. Bell and C. Volinsky, “Matrix factorization techniques for recommender systems,” *Journal*

Copyright © by ORSJ. Unauthorized reproduction of this article is prohibited.

Computer, **42**(8), pp. 30–37, 2009.
[10] T. Mikolov, K. Chen, G. Corrado and J. Dean, "Efficient estimation of word representations in vector space," arXiv: 1301.3781, 2013.
[11] Q. Le and T. Mikolov, "Distributed representations of sentences and documents," In *Proceedings of the 31st Intarnational Conference on Machine Learning*, pp. 1188–1196, 2014.
[12] V.-T. Phi, L. Chen and Y. Hirate, "Distributed representation-based recommender systems in E-commerce," 第 8 回データ工学と情報マネージメントに関するフォーラム，C8-1, 2016.
[13] P. Loyola, C. Liu and Y. Hirate, "Modeling user session and intent with an attention-based encoder-decoder architecture," In *Proceedings of 11th ACM Conference on Recommender Systems*, pp. 147–151, 2017.
[14] T. Li, Y. Ma, J. Xu, B. Stenger, C. Liu and Y. Hirate, "Deep heterogeneous autoencoders for collaborative filtering," In *Proceedings of International Conference on Data Mining*, pp. 1164–1169, 2018.

 Copyright © by ORSJ. Unauthorized reproduction of this article is prohibited.