複写機の定着装置における交換時期の推定

中田　和秀

キーワード：異常検知，判別問題，機械学習

本稿は，谷川 奈穂さんによる 2015 年度東京工業大学 修士論文をもとに加筆修正しました．

1. はじめに

近年，さまざまな分野において，遠隔監視技術を活用した予防保守[1]の考え方が浸透してきている．富士ゼロックス株式会社では，この遠隔監視で得たデータを活用して，より付加価値の高いサービスを顧客に提供することを模索している．その一例として，トラブルを事前に予測するトラブル予兆検知がある．本研究では，トラブル予兆検知における取り組みの一つである「定着装置の交換予測」という問題を扱う．複写機で使用されている定着装置は，用紙の上に張り付いたトナーを用紙に定着させるためのもので，サービスエンジニアが規定の基準に従って交換している．定着装置の部材は高価な消耗品であり，交換費用は業者がすべて負担しているため，不必要な交換は業者に損失をもたらす．また，定着装置が適正な時期に交換されていないものは，定着装置の劣化に伴って複写機が急停止してしまう可能性がある．蓄積された複写機の稼働情報を活用して，定着装置の交換の必要性を判断するための判別モデルを作成することにより，不必要な交換や未交換による急停止を防止することが期待できる．

2. 問題設定

2.1 稼働情報

分析に使用するデータは 2012 年 2 月 20 日から 2015 年 1 月 28 日に取得された 60,778 台の複写機に関する稼働情報である．データは延べ 4,041,894 件あり，一つのデータは Daily 情報・Fault 情報・1 万枚印刷ごとの情報・訪問情報・交換情報など 161 項目からなる．

なかた　かずひで
東京工業大学 工学院
〒 152–8552 東京都目黒区大岡山 2–12–1 W9–60
nakata.k.ac@m.titech.ac.jp

表 1　評価指標に用いるクロス表

		判別したクラス		
		継続	交換	不明
実際の	継続	継続正答数	誤抽出数	不明数 1
クラス	交換	取りこぼし数	交換正答数	不明数 2

2.2 判別問題と評価指標

本研究では現場からの要請を踏まえ，データ取得日から 3 カ月以内における定着装置の交換の必要性を予測する．ただし，より信頼性の高い結果を得るために，定着装置を「交換」するか，「継続」するかだけでなく，「不明」という判断も含めて判別することにする．判別モデルの評価指標として，表 1 の分類に基づいた評価指標 1, 2 を導入している．なお，各番号は優先順位を意味する．したがって，Replace 率と Unknown 率が基準を満たしたうえで，重み和が最小となる判別を最も性能がよいモデルとみなす．また，本研究における不必要な交換は「誤抽出数」，未交換による急停止は「取りこぼし数」に対応している．今回は，不必要な交換をより防止したいため，以下のような重み和を用いている．

1. Replace 率が 0.35〜1.4%で，Unknown 率が 30%以下

$$\text{Replace 率} = \frac{\text{誤抽出数} + \text{交換正答数}}{\text{検証データ数}}$$

$$\text{Unknown 率} = \frac{\text{不明数 1} + \text{不明数 2}}{\text{検証データ数}}$$

2. 重み和が最小

$$\text{重み和} = 126 \times \text{誤抽出数} + 1 \times \text{取りこぼし数} + 0.1 \times (\text{不明数 1} + \text{不明数 2})$$

3. 提案手法

3.1 前処理

業務中に収集した生データは，そのままでは分析に不適当であることも多い．本データも分析の精度を上げるため，まず次で述べるようなデータの前処理を行った．

[1] 予防保守とは，不具合を顕在化させ，未然に障害の発生を防ぐための考え方や対応を指す．

・異常値を含む・欠損値であるなど判別に適していない複写機 5,900 台と，その複写機のデータ 405,613 件を除外した．
・データの各項目は異なる性質を持っている．特徴量として使用できる項目を選択し，各項目の性質に合わせてダミー変数化，対数変換，および標準化を施した．
・定着装置の劣化理由として，短期間における複写機の過度な使用が考えられる．そこで，各データの特徴量に対して，前回に格納された同じ複写機のデータの特徴量との変化量を算出し，新しい特徴量として導入した．

3.2 3 クラス判別

本研究では，交換／継続の 2 クラスが付与された過去データ（学習データ）で学習を行い，現在のデータに対し交換／継続／不明の 3 クラスに判別することになる．このような 3 クラス判別の関連研究として Ternary Spam Filtering [1] がある．しかし，本研究とは評価指標が異なるため，二つの閾値を用いて 3 クラスに判別する点のみを参考にした．提案手法では，判別モデルによって計算された交換の必要性を意味する予測値 f と，閾値 α, β を用い，以下のように判別を行う．

$$\begin{aligned} f < \alpha &\quad : \text{継続と判別} \\ \alpha \leq f < \beta &\quad : \text{不明と判別} \\ \beta \leq f &\quad : \text{交換と判別} \end{aligned}$$

3.3 判別モデル

定着装置は頻繁に交換されないため，3 カ月以内に交換される状況でのデータが極端に少ない．このように，データ中のクラス割合が極端に多かったり，少なかったりするアンバランスなデータを不均衡データという．不均衡データを使って学習した判別モデルは，多数クラスのデータに対しては性能がよいが，少数クラスのデータに対しては性能が悪い傾向にあることが知られている．そこで本研究では，サンプリング手法の一つであるアンダーサンプリングを使用して新たに学習データを作成し，クラス割合を平準化した．また，Boosted Random Forest [2] に AdaC1 [3] という不均衡データを考慮しているアンサンブル学習を適用し，誤抽出数と取りこぼし数を抑える判別モデルの作成を可能とした．

3.4 閾値

学習データに付与している交換／継続の情報は，複写機の 3 カ月以内の急停止の有無を意味する．そのため，学習データは少なくとも 3 カ月以上前に取得されたデータでなければならない．しかし，そのようなデータで学習した閾値を使うと，評価指標 1 の Replace 率と Unknown 率が規定の範囲から大きく外れる結果となった．そこで，閾値は 1 週間前から前日までに取得した直近データを使って学習することにした．ただし，そのデータには本来学習に必要な交換／継続の情報が付与されていないため，直近データにおいて Replace 率が 0.35%，Unknown 率が 30%（評価指標 1 の下限値と上限値）となる α, β を閾値とした．これにより，妥当な Replace 率と Unknown 率をもつ判別結果を得ることが可能となった．

表 2 検証期間の判別結果

		判別したクラス		
		継続	交換	不明
実際の	継続	484,115	2,328	205,158
クラス	交換	2,996	87	2,744

4. 数値実験

提案手法の性能検証のため，全データを学習期間と検証期間に分け，学習期間のデータで判別モデルのパラメタと閾値の学習を行い，検証期間で提案モデルの性能を検証した．表 2 がその結果である．Replace 率は 0.35%，Unknown 率は 29.8%と評価基準の範囲にほぼ収まっている．しかし，「交換」の予測を外しているものも多く，実務で利用するにはまだ多くの改善が必要である．複写機は数十年にわたって品質改善が重ねられてきた製品であり，明快で単純な理由での定着装置の異常はあまり起こらなくなっている．そのため，現状取られているデータだけでは異常予知の精度を上げることは難しく，別の角度からの分析が必要である可能性がある．

謝辞 富士ゼロックス株式会社からは，研究に関する多くの示唆をいただき，数値実験に必要なデータの提供を受けました．心より感謝いたします．

参考文献

[1] W. Zhao and Z. Zhang, “An email classifiation model based on rough set theory,” In *Proceedings of the 2005 International Conference on Active Media Technology(AMT05)*, pp. 403–408, 2005.
[2] Y. Mishina, M. Tsuchiya and H. Fujiyoshi, “Boosted Random Forest,” In *Proceeding of the International Conference in Computer Vision Theory and Applications(ICCVTA)*, pp. 594–598, 2014.
[3] Y. Sun, M. S. Kamel, A. K. Wong and Y. Wang, “Cost-senstive boosting for classifiation of imbalanced-data,” *Pattern Recognition*, **40**, pp. 18–36, 2007.