Vol.2020-MPS-131 No.8
2020/12/17

顧客分類のための多次元時系列データに基づく特徴量選択法

雲居 玄道[1,a)]　後藤 正幸[1]

概要：近年，ネットワークに接続された電子機器から，その利用履歴である多次元時系列データが収集可能となっている．電子機器メーカーの恒常的なマーケティング活動において，顧客であるユーザの電子機器の利用履歴データを詳細に分析することは多大な価値を生むことが期待できる．しかし，取得された膨大なデータを全て人手で分析することは，計算量や人的コストの点から現実的ではない．そのため，一般的に経験に基づき注目する変数を絞り込み，モニタリングすべき特徴量をリスト化しておくことは，実務上のメリットがある．このような，膨大なデータから注目すべき特徴量の生成法は特徴量選択と呼ばれ，何かしらの合理的な基準によって特徴量を決定することが望ましい．本研究では，対象となる電子機器の利用がカートリッジの純正・非純正の選択要因に注目する．その上で，遺伝的アルゴリズムに基づく特徴量選択法を提案する．これにより恒常的な運用が可能な特徴量を同定が可能になることを示す．

キーワード：特徴量選択，Random Forest，多次元時系列データ，遺伝的アルゴリズム

Feature Selection Method Based on Multidimensional Time Series Data for Customer Classification

Gendo Kumoi[1,a)]　Masayuki Goto[1]

Abstract: Recently, it has become possible to collect multidimensional time series data, which is the usage history of electronic devices connected to the Internet. In the constant marketing activities of electronics manufacturers, detailed analysis of the usage data of customers' electronic devices is expected to be of great value. However, it is not reasonable to analyze all the huge amount of data acquired manually due to the computational complexity and human cost. Therefore, there are practical benefits to narrowing down the variables of interest, generally based on experience, and keeping a list of features to monitor. Such a method of generating noteworthy features from large amounts of data is called feature selection, and it is desirable to determine the features by some proper criteria. In this study, the use of the target electronics is a factor in the selection of genuine and non-genuine cartridges. We then propose a feature selection method based on a genetic algorithm. We show that the proposed method allows for the identification of features that can be used permanently.

Keywords: Feature Selection, Random Forest, Time Series Data, Genetic Algorithms

1. はじめに

近年，Internet of Things（以下，IoT）社会の実現に向かい，関連技術が発達し，様々な電子機器がネットワークに接続されるようになっている．これに伴い，ユーザが外出先から遠隔操作で家庭の家電を操作できる [1] など，新たなサービスの享受が可能となり利便性が向上している．また，これらの電子機器や運用サービスを提供する企業においても，ユーザの許諾を得る形で，機器の利用履歴の収集が可能となっている．これらの利用履歴データは継続的に測定されるビックデータであり，その貴重なデータを有効活用し，適切に分析した上で，企業活動へフィードバッ

[1] 早稲田大学
Waseda University, Shinjuku, Tokyo 169–8555, Japan.
[a)] moto-aries@ruri.waseda.jp

クする方法を構築することが重要な課題となっている [2].

この技術の先駆的な機器の 1 つにプリンター製品がある．プリンターは，近年，PC へ直接接続する形からネットワークを介した接続へと変わりつつあり，家庭でもネットワークに接続される機器となっている．このため，これらの機器を提供する企業は，データ提供に協力を承諾したユーザから日々の利用履歴を収集することが可能となる．利用履歴は，印刷枚数や消耗品の交換回数，ジャムなどのエラー発生数など多岐にわたる変数が時系列データとして蓄積される．こうしたデータは多次元時系列データと呼ばれる．これらのプリンターの利用履歴データは変数が多いため大変リッチな時系列のデータである．そのため，どのように分析し，どのような目的に活用できるのかについての検討は未発達である．

一般に，プリンターを購入して利用しているユーザには，様々なタイプの顧客が存在している．そのため，これらを分類整理しておくことはマーケティング活動の面から有用である．特にプリンターの利用履歴においては，顧客の利用特性に関するデータとして，使用しているカートリッジの種類がとても重要な情報となっている．このカートリッジの種類とは，機器を提供する企業から販売されている純正品と他の企業から販売されている非純正品のことを示す．非純正品の使用は，顧客がランニングコストを抑える目的で使用されることも多い．一方で，用紙のジャムなどの発生率も上昇するとされている．機器を提供する企業にとって，顧客が純正品・非純正品のどちらを使用するかは，売上のみならず，機器の満足度にも直結するため，重要な関心事となっている．このような観点から，収集され蓄積されたデータから顧客の利用特性に基づいて，カートリッジ選択の要因を分析をすることは，とても重要なテーマである．

一方で，これらの膨大な多次元時系列データから，どのように特徴量を構成し，顧客分析に用いるかについては，分析者の経験や技能的なスキルに依存していることが多い．印刷枚数や消耗品の交換回数，エラー発生数など多岐にわたる時系列データから，どのように特徴量を構成すれば顧客理解に結び付くのかは，マーケティング分析の観点からも大変重要である．特に現場レベルでの日々のマーケティング活動を考慮すると，データ取得や分析コストを抑えるために，特徴量の構成は小規模なものが望ましい．そこで本研究においては，顧客の利用特性を特徴づける特徴量を機械学習を用いて自動的に選択する方法を提案する．

まず，収集されるデータは，印刷枚数や消耗品の交換回数，エラー発生数などの時系列データという多次元時系列データとなっている．これらのデータから特徴量を選択する際には，変数，期間，合成変数の 3 つの視点がある．変数は，印刷枚数や消耗品の交換回数などを意味する．期間は，得られた時系列データに対して，対象とするデータの時間を意味する．これらの変数と期間が定まったとしても，時系列データに対して，その統計的特徴を捉えるための統計量として，最大値や最小値，平均値や分散などを用いることも考えられ，これらが合成変数である．

これら 3 つの視点に基づき，顧客の利用特性としてデータより得ることのできる顧客の使用するカートリッジの種類に対する分類器を学習することを提案する．その上で，分類器へ入力する特徴量については，想定される様々な特徴量から分類器の性能の良い特徴量を探索することで，顧客の利用特性を決定づける要因を抽出することを考える．

一方で，候補となる特徴量の種類は極めて膨大となってしまうため，全ての組合せを考慮して比較検討することは現実的ではない．そこで，遺伝的アルゴリズム [3], [4] を用いて，効率的に分類に寄与する特徴量を探索し，特徴量選択を行う方法を提案する．これらの提案手法を実データに適用し，提案手法の有効性を示すと共に考察を与える．得られる結果から，顧客が使用するプリンタの使用実績として適切な特徴量が得られ，興味深い考察が得られることを示す．得られる特徴量は，単に顧客の分類に適切な特徴空間の構成に利用されるだけでなく，実時間のモニタリングにおいて注目すべき統計量を与えてくれるという効用がある．

2. 関連研究

特徴量選択の研究は機械学習への入力を構成する技術として発展してきた．特徴量選択の 1 つに，機械学習の精度を向上させる変数を探索的に選択する手法がある．これらの多くは，多次元データに対する選択すべき変数の探索に Genetic Algorithms（以下，GA） を用いた手法 [5], [6], [7], [8], [9] がある．

次に，特徴量選択として，どのような期間に着目するべきかという課題は時系列予測の観点から研究されている．特に株価などの時系列データ未来に上昇するか下降するかは，重要なテーマである．この上下を予測する分類器の入力として，選択すべき期間を探索する研究 [10] がある．

加えて，これらの時系列データから合成変数法であるテクニカル指標から，どの特徴量を選択するかという研究 [11], [12] がある．

3. 提案手法

3.1 着眼点

一般的な特徴量選択の手法は，与えられた特徴の集合を元に，その次元を削減することを目的として行われている．これらは，膨大な入力データの次元から分析に有用な特徴量を選択し，特徴空間の次元を適切なレベルまで落とすことを意味している．そのため，通常は各特徴量の重要度を何らかの方法で計算し，特徴量を選択している．

本研究の対象としている多次元時系列データにおいて

は，印刷枚数やエラーの発生回数などの複数の値が観測されており，これらより分析に有用な特徴量を選択することは有用であると考えられる．一方で，これらが定期的に観測される時系列データであることから，特徴量としては各時点におけるデータも有している．この生データをそのまま分析に用いることも考えられるが，一定期間に対する最小値や最大値，平均値や分散など，様々なデータの加工によって与えられる統計量を特徴量として用いる方法も考えられる．このような考えられる様々な特徴量に対し，どういった特徴量を選択することが有用であるかは，経験則のみでしか判断されていなかった．

そこで，本研究においては，多次元時系列データから顧客の利用特性を分ける要因となる特徴量を探索的に求める手法を提案する．

3.2 提案手法

本研究においては，多次元時系列データから顧客の利用特性を分ける要因となる特徴量を選択する手法を提案する．

対象となるデータには，顧客の利用特性として，使用しているトナーやインクが純正品・非純正品に関するフラグが存在している．トナーやインクについて，純正品・非純正品のうち，どちらを使用するかは，顧客を理解する上で大変重要な指標となっており，この差異を生む要因を特定することは，ビジネス上，非常に価値がある．本研究では，この点に着目し，顧客を“純正品の使用”と“非純正品の使用”の 2 カテゴリに分ける二値分類器を学習することを考える．その上で，この二値分類器を構成する際に，有用となる特徴量を探索的に求めることにより，経験則のみでしか判断されて来なかった，顧客の利用特性を分ける特徴を明らかにする．

このとき，Junior ら [12] のように対象とする期間を合成変数ごとに定めていた．しかし，本研究の対象が多次元データであることから取得・分析コストを考えるならば，期間は一定の方が望ましい．そこで，本研究では次元，期間，合成変数の種類の 3 つの視点に対して，それぞれ遺伝子をもつ GA を考える．このとき，合成変数の種類に関する遺伝子は，選択された次元に依存して存在する．そのため，全ての遺伝子について検討する必要はない．そこで，選択された次元に対して GA を適用することを考える．そのため，本提案手法は GA の二重構造から成り立つ．

ここで，GA を用いた特徴量探索において，良い遺伝子であるかどうかを示す値を適応度と呼ぶ．本研究では，第一に用いる分類器の性能が重要となる．そこで，ランダムフォレスト [13] を用いて，顧客のフラグを用いた二値分類器を学習し，その Area Under the Curve（以下，AUC）[14] を適応度として用いる．ここで，AUC とは Receiver Operating Characteristic 曲線下の面積のことである．この AUC は分類器の性能を表し，完全な分類を達成するときには “1”．ランダムな分類のときには “0.5” をとる．また，学習データへの当てはまりではなく汎化能力の高い分類器を構築すべきことから，この AUC には 10 分割交差検証を行った平均値を用いる．

これらの探索により，顧客の利用特性を分けるのに有効な次元，期間，合成変数が同定可能となる．そのため，これらの結果を元に詳細な分析を行うことで，マーケティング施策の立案が可能となる．

4. 実データ分析

本章では，提案する手法の有効性を検証するため，実際にネットワークを介して収集されたプリンターの利用履歴実データを分析し，その結果を検証する．

4.1 分析対象データ

本研究では，同ドメインの事業を展開する A 社から提供を受けた，2017 年 12 月〜2018 年 2 月の期間に毎週収集された 12 週間の週次利用履歴データを対象とする．また，顧客のラベルは 2018 年 5 月末のものを対象とする．ただし，プリンターはユーザによって電源がシャットダウンされていたり，ネットワークから切り離されていたりする場合，利用履歴データを取得することができず，いわゆる欠損が生じる．ここでは，対象とした期間において，欠けることなくデータを送信された 570 台のプリンターを分析対象とする．

4.2 特徴量

本研究で対象とするデータは，週間印刷枚数やジャム回数やインク残量など 121 変数である．期間は，実験対象データの 12 週間のうち，週毎に選択を行う．合成変数は，一般的なものとして，最大値，最小値，平均値，中央値，分散の 5 種を対象とする．

4.3 実験条件

これらの特徴量を元に表 1 に示した比較手法および提案手法を比較する．比較手法 1,2 は，単純な手法としての比較である．また，比較手法 3 は，Junior ら [12] の手法を遺伝子を 2 つもつ GA による特徴量選択手法として捉え直し，本研究に適用したものである．

本研究において，使用するデータは二値分類において不均衡データとなっている．そのため，オーバーサンプリング [15] を行う．性能評価には，適応度として使用した AUC を用いる．また，遺伝的アルゴリズムにおけるパラメータは表 2，分類器であるランダムフォレストのパラメータは表 3 のように設定した．

4.4 実験結果

実験結果を以下の表 4 に示す．比較手法 3 および提案手

表 1 比較手法
Table 1 Comparison method

手法名	詳細
比較手法 1	全期間の全変数のデータを入力（121 × 12 種）
比較手法 2	全期間の全変数の合成変数を入力（121 × 5 種）
比較手法 3	期間，全変数の合成変数の 2 種の遺伝子における GA

表 2 遺伝的アルゴリズムのパラメータ
Table 2 Parameters of the genetic algorithm

個体数	300
交差確率	0.50
突然変異確率	0.20
繰り返し回数	50

表 3 Random Forest のパラメータ
Table 3 Parameters of Random Forest

木の数	100
基準	Gini 係数
木の最大特徴量	$\lfloor\sqrt{\text{特徴量数}}\rfloor$

法で選択された期間を 5 に示す．提案手法により選択された変数，合成変数の詳細を表 6,7 に示す．

表 4 実験結果
Table 4 Experimental results

	変数の数	総特徴量数	AUC
比較手法 1	121	1,452	0.553
比較手法 2	121	605	0.680
比較手法 3	119	314	0.736
提案手法	**45**	**111**	**0.760**

4.5 実験結果の解釈と考察

表 4 より，変数，総特徴量数共に最も少ない結果となった．GA での探索を行う際に，提案手法では特徴量ごとに選択するか否かの遺伝子を用意した．結果として，比較手法 3 と比べて少ない変数となったことが考えられる．その上，AUC も最も高い値を示した．比較手法 3 においても，提案手法によって得られた解は，探索可能な範囲に存在する．しかし，全変数の全合成変数に対する遺伝子のため，本実験では提案手法の解を得られなかったと考えられる．これらの結果より選択する変数，期間，合成変数に対してそれぞれ遺伝子を用意する提案手法は，効率的な探索が行えることが分かる．その上，AUC の上昇が見られたことから，より顧客の利用特性を分ける特徴量を選択できたと考えられる．

表 5 に，選択された期間を示した．Junior ら [12] らは，合成変数ごとに期間を定めていた．これに対し，本研究で対象とする問題では，全変数で期間は一定なことが望ましい．そこで，比較手法 3 も提案手法と同様に期間については，全変数で 1 つを選択する手法とした．この結果，比較手法，提案手法ともに 2018 年 2 月は 4 週間全てが選択されている．これは目的変数までの時差が少ないデータが望まれるという結果である．このことから，時系列予測として，一般的な主張と同様な期間選択が行われている．そのため，適切な期間選択ができたと考えられる．また，提案手法では，2017 年 12 月，2018 年 1 月では月初の期間が選択された．これにより，過去のデータの必要性が示されたと考えられる．最後に，比較手法は提案手法よりも短い期間の選択となった．しかし，表 4 より，比較手法 3 は一定の精度を保っている．このことから，変数を多く取れば期間を短く，変数が少なくなれば期間が長く必要となる可能性が考えられる．

表 6 に，選択された変数の種類を示した．最も多いものは印刷枚数である．これは，印刷する紙の種類や印刷に使用した機器など様々な種類別にデータが蓄積され，全変数の半数近くを占めている．このことから，印刷枚数が重視される結果となったと考えられる．また，このように変数間は相関が強いものが多くある．例えば，一般的に残量警告が発生し，カートリッジ交換が行われる．そのため相関の強いものとして，例えばカートリッジ交換回数と残量警告がある．これらについて，選択された変数をみると，黒のカートリッジ交換回数は選択されず，残量警告は選択された．このように，相関の強い変数はどちらか 1 つが存在すればよく，提案手法は選択ができたと考えられる．

表 7 に，選択された合成変数ごとに変数の数を示した．選択された変数が 45 であることから，それぞれの合成変数として約半数が選ばれていることがわかる．特に分散が必要とする変数が多かった．これは，日々の印刷におけるバラツキが重要であるということが示唆される．

4.6 顧客の利用特性分析

実験結果を踏まえて，層別分析を行った．層別には実験データで使用した顧客ラベルを用いて，2018 年 5 月末時点での純正品を使用する顧客（以下，純正顧客）・非純正品を使用する顧客（以下，非純正顧客）に分ける．層別されたデータに対して，選択された期間，変数に基づき，合成変数ごとに平均値を算出した．そして，層間の差が大きい上位 5 変数を表 8–12 に示す．

具体的には，表 8 は，表 7 の最小値が選択された 22 変数について，表 5 の期間の最小値を算出する．次に，純正・

表 5 選択された期間
Table 5 Selected time period

手法	2017 年 12 月				2018 年 1 月				2018 年 2 月			
	1	2	3	4	5	6	7	8	9	10	11	12
比較手法 3				■	■				■	■	■	■
提案手法	■	■		■	■	■			■	■	■	■

表 6 選択された変数
Table 6 Selected Dimensions

種類名	変数の数
印刷枚数	25
カートリッジ	7
試行回数	5
警告回数	3
ジャム回数	2
スキャン	2
通電時間	1

表 7 選択された合成変数
Table 7 Selected Dimensions

	最小値	最大値	平均	中央値	分散
変数の数	22	22	22	19	26

表 8 純正・非純正顧客間の差（最小値）
Table 8 Difference between categories (Min)

No.	変数	大小関係
1	通電時間	純正 > 非純正
2	モノクロ. 印刷枚数	純正 < 非純正
3	PC. 印刷枚数	純正 < 非純正
4	スキャン数	純正 > 非純正
5	モノクロ-両面. 印刷枚数	純正 < 非純正

表 9 純正・非純正顧客間の差（最大値）
Table 9 Difference between categories (Max)

No.	変数	大小関係
1	黒カートリッジ. エラー	純正 < 非純正
2	A4. 印刷枚数	純正 < 非純正
3	PC-両面. 印刷枚数	純正 < 非純正
4	黒. カートリッジ残量警告	純正 < 非純正
5	マゼンタ. カートリッジ使用量	純正 < 非純正

表 10 純正・非純正顧客間の差（平均値）
Table 10 Difference between categories (Average)

No.	変数	大小関係
1	PC-両面. 印刷枚数	純正 < 非純正
2	A4. 印刷枚数	純正 < 非純正
3	PC. 印刷枚数	純正 < 非純正
4	黒カートリッジ. エラー	純正 < 非純正
5	モノクロ-両面. 印刷枚数	純正 < 非純正

表 11 純正・非純正顧客間の差（中央値）
Table 11 Difference between categories (Median)

No.	変数	大小関係
1	A4. 印刷枚数	純正 < 非純正
2	PC-カラー. 印刷枚数	純正 > 非純正
3	通電時間	純正 > 非純正
4	シアンカートリッジ. エラー	純正 > 非純正
5	マゼンタ. カートリッジ使用量	純正 < 非純正

表 12 純正・非純正顧客間の差（分散）
Table 12 Difference between categories (Variance)

No.	変数	大小関係
1	A4. 印刷枚数	純正 < 非純正
2	PC-カラー. 印刷枚数	純正 > 非純正
3	PC-両面. 印刷枚数	純正 < 非純正
4	両面. 印刷枚数	純正 < 非純正
5	シアン. カートリッジ残量警告	純正 < 非純正

非純正顧客それぞれで平均値を取る．最後に，純正顧客 – 非純正顧客と差を取り，その絶対値が大きいものから順に並べた．

このため，表 8–12 の差は，正の値は純正顧客が大きいことを示し，負の値は非純正顧客が大きいことを示している．

表 8–11 において，印刷枚数の多くが負の値を取ることから非純正顧客で多いことが分かる．特に印刷枚数の多い典型の A4 では，最大値，平均値，中央値ともに上位にあり，顧客の利用特性を決定づける要因と分かる．一方で，PC からのカラー印刷枚数は，純正顧客の方が多い．そして，黒カートリッジ残量警告やモノクロ印刷枚数は非純正顧客が多い．これらのことから，モノクロ印刷が多い顧客は非純正顧客に変化しやすく，カラー印刷が多い顧客は純正顧客に留まりやすいことが分かる．

表 12 で，A4 印刷枚数の分散は非純正顧客が大きくなっている．本実験で使用したデータは週次のデータである．そのため，分散が大きいとは，週毎に印刷枚数の変動が大きいことを意味する．つまり，非純正顧客は月初や月末など週による偏りがある顧客である．

表 8 では，通電時間が非純正顧客が短いという結果になった．これが最小値であることを考えると，どれくらい機器の電源を切っているかということになる．本来であれば，印刷枚数の多い非純正顧客は通電時間も長くなると想定できる．しかし，表 11 の中央値でも非純正顧客が短く

なっている．さきほどの分散と合わせて考えると，非純正顧客は使用時期に偏りがある．そのため，使用する時に電源を入れるといった運用を行っている可能性がある．

また，カートリッジエラーも非純特定の方が多いことがわかる．これは，非純正品を使用することにより，増加するものと考えられる．

5. 考察

本研究では，顧客の利用特性を分ける分類器を構築し，このとき入力とする特徴量を，GA を用いた探索を行った．この際，印刷枚数や通電時間といった変数だけではなく，期間や合成変数の組合せで特徴量選択を行った．これにより，精度の高い分類器が構築された．すなわち，提案手法は現場レベルで解釈可能な低コストかつ精度の高い分類器を構築できる手法であることが分かる．

その上で，選択された特徴量をもとに，顧客の利用特性を決定づける要因について詳細な分析を与えた．従来は，多数のユーザの多次元時系列データは膨大なデータであるため，期間や対象とする変数は経験に依るものであった．これに対して，提案手法は，対象とした顧客の利用特性であるラベルに対応した特徴量選択である．この結果を用いて分析することは，顧客の利用特性を決定づける要因が自動的に選択されることになる．

例えば，分析より通電時間が短く大量の印刷を行う顧客は非純正顧客となりやすいと分かった．これは，顧客のコスト意識に起因するものと捉えることができる．そのため，頻繁に電源を切るなど，コスト削減に対する意識の強い顧客に対してのターゲット・マーケティングなどの施策立案などが期待される．

一方で，提案手法のみでは因果の関係性は明らかではないことは注意が必要である．

6. まとめと今後の課題

提案手法により，カートリッジの純正品と非純正品の分類問題に対して精度の高い分類器を構築することができた．これにより，変数，期間，合成変数の遺伝子をもつ GA は，有効な解を実時間で探索できる優れた手法であると考えられる．これまで，Random Forest をはじめ，予測精度が高いとされる分類器は，マーケティング活動へ適用する際の解釈が難しいといった問題があった．これに対して，本手法の探索結果を用いることで，変数ごとの顧客の利用特性に関する解釈が可能となった．これにより，マーケティング活動への活用についても一定の成果を与えた．

本手法では GA の適応度として，分類器の 10 分割交差検証の結果を用いた．分類器の評価として適切である一方，計算時間が多くかかるという問題がある．そのため，計算量の削減が 1 つの課題である．また，実ビジネスにおいては，因果関係を知ることが大切である．しかし，提案手法では，そこまでを明らかにはできず，分析者に委ねることになる．このため，因果推論モデルなどによる因果関係の同定や合成変数として複数の変数を組合せた合成変数の自動生成などが今後課題である．

謝辞

本研究を進めるにあたり，貴重な実データを提供頂いた A 社の皆様に深く感謝いたします．

参考文献

[1] 矢崎孝一，伊藤栄信，坂本拓也，二村和明，Others：スマホ認証を用いた IoT 機器サービスの簡易利用方式，研究報告マルチメディア通信と分散処理 (DPS)，Vol. 2017, No. 5, pp. 1–8 (2017).
[2] 村上憲郎：SNS と IoT（Internet of Things）が切り拓く，ビッグデータ 2.0 の世界，情報管理， Vol. 56, No. 2, pp. 71–77 (2013).
[3] Fogel, D. B.: *Evolutionary Computation: The Fossil Record*, Wiley-IEEE Press, 1st edition (1998).
[4] 北野宏明，Others：遺伝的アルゴリズム，人工知能学会誌， Vol. 7, No. 1, pp. 26–37 (1992).
[5] Goldberg and E, D.: Genetic Algorithms in Search, *Optimization, and MachineLearning* (1989).
[6] Goldberg, D. E., David Edward, G., Goldberg, D. E. G. and Visiting Assistant Professor of History David E Goldberg: *Genetic Algorithms in Search, Optimization, and Machine Learning*, Addison-Wesley Publishing Company (1989).
[7] Krishna, K. and Narasimha Murty, M.: Genetic K-means Algorithm, *IEEE Trans. Syst. Man Cybern. B Cybern.*, Vol. 29, No. 3, pp. 433–439 (1999).
[8] Kim, Y., Street, W. N., Russell, G. J. and Menczer, F.: Customer Targeting: A Neural Network Approach Guided by Genetic Algorithms, *Manage. Sci.*, Vol. 51, No. 2, pp. 264–276 (2005).
[9] Frohlich, H., Chapelle, O. and Scholkopf, B.: Feature Selection for Support Vector Machines by Means of Genetic Algorithm, *Proceedings. 15th IEEE International Conference on Tools with Artificial Intelligence*, ieeexplore.ieee.org, pp. 142–148 (2003).
[10] Gan, M., Cheng, Y., Liu, K. and Zhang, G.-L.: Seasonal and Trend Time Series Forecasting Based on a Quasi-Linear Autoregressive Model, *Appl. Soft Comput.*, Vol. 24, pp. 13–18 (2014).
[11] Kara, Y., Acar Boyacioglu, M. and Baykan, Ö. K.: Predicting Direction of Stock Price Index Movement Using Artificial Neural Networks and Support Vector Nachines: The Sample of the Istanbul Stock Exchange, *Expert Syst. Appl.*, Vol. 38, No. 5, pp. 5311–5319 (2011).
[12] Junior, N. R. and Nievola, J. C.: A Generalized Financial Time Series Forecasting Model Based on Automatic Feature Engineering Using Genetic Algorithms and Support Vector Machine (2018).
[13] Breiman, L.: Random Forests, *Machine learning*, Vol. 45, No. 1, pp. 5–32 (2001).
[14] Fawcett, T.: An Introduction to ROC Analysis, *Pattern Recognit. Lett.*, Vol. 27, No. 8, pp. 861–874 (2006).
[15] Chawla, N. V., Bowyer, K. W., Hall, L. O. and Kegelmeyer, W. P.: SMOTE: Synthetic Minority Over-Sampling Technique, *J. Artif. Intell. Res.*, Vol. 16, pp. 321–357 (2002).