情報処理学会論文誌　コンシューマ・デバイス & システム　Vol.6　No.2　14–22 (Sep. 2016)

コンシューマ・システム論文

ホームネットワーク内接続機器の情報を活用した世帯人数推定システム

美原 義行[1,a)]　山口 徹也[1]　高倉 健[2]

受付日 2016年2月28日, 採録日 2016年7月7日

概要：本論文では，ホームネットワークに接続された機器の情報を活用して，機器を所有している世帯の人数を推定するシステムについて述べる．機器から取得可能な情報とは，機器の種別やメーカ名等の機器名情報と，各機器の利用状態を記録した利用情報のことである．本システムでは，膨大な利用情報に対して，世帯の特徴を表現できるよう丸め処理を実施し，学習して推定モデルを作成することで世帯人数を推定する．世帯人数を把握できることにより，レコメンドに向けある商品やサービスに適した世帯を抽出することができ，また，世帯ごとの活動を測定するマーケティングにも応用することが可能となる．本システムを実ホームネットワークに適用し，各ホームネットワークから機器名情報と利用情報を収集でき，これらの情報から世帯人数推定のモデルを構築できることを確認した．本システムによる世帯人数推定の精度評価として，1,000 世帯からアンケートにより機器名情報と利用情報，世帯人数情報を収集し，世帯人数の推定を実施した．その結果，ある特定の目的変数とそれ以外の目的変数のどちらに適合するかを判定する二値分類にて平均 83.7%の精度で推定できた．そして，商品やサービスのレコメンドに適した世帯である，可処分所得が多い 1 人世帯においては 89.5%の適合率で推定でき，子どもを含む世帯が多い 3 人以上世帯においては 88.3%の適合率で推定できた．機器に関する情報から，高い精度で世帯の人数を推定でき，世帯人数推定における機器に関する情報の有効性を確認することができた．

キーワード：ホームネットワーク，機械学習，情報家電，ネットワークプロトコル

A System That Estimates a Household Size Using Information from Devices Connected to Home Network

Yoshiyuki Mihara[1,a)]　Tetsuya Yamaguchi[1]　Takeshi Takakura[2]

Received: February 28, 2016, Accepted: July 7, 2016

Abstract: In this article, we propose a system that estimates household size from the information held by devices connected to the home network. This information includes device name information, such as device type or manufacturer name, and usage status information of each device. This system agglomerates the vast amount of usage status information in order to estimate the feature of each family. This system estimates the household size by learning device information and by making a learning model. Estimating the household size allows us to provide services that more appropriate to the family, or sell the information to other marketing services. We verify that the system collects device name information and usage status information from actual home network and makes the learning model. We submit questionnaires to 1,000 families to gather household size, device names, and the usage status information and then process the data to extract household size. Our system estimates the household size with 83.7% precision for binary classification. 89.5% precision for one-person household, which has a lot of disposal income and is appropriate for recommendation of service or products. 88.3% precision for more than three household, in which there are many families with children. Clearly the proposed system can estimate household size with a high degree of accuracy. The results show the effectiveness of using device information to estimate the household size. Moreover, we verify that using agglomerated usage status yields higher accuracy than if it is not used. We can the effectiveness of round usage status information.

Keywords: home networks, machine learning, networked appliances, network protocols

[1] 日本電信電話株式会社 NTT サービスエボリューション研究所
NTT Service Evolution Laboratories, Yokosuka, Kanagawa 239–0847, Japan
[2] 日本電信電話株式会社 NTT サービスイノベーション総合研究所
NTT Service Innovation Laboratory Group, Yokosuka, Kanagawa 239–0847, Japan
[a)] mihara.yoshiyuki@lab.ntt.co.jp

1. はじめに

昨今，パーソナルコンピュータだけでなく，スマートフォンやゲーム機，センサ等，ホームネットワークに接続可能な機器が増加してきており，それら機器を利用したサービスが普及しつつある．さらに，それらの機器を所有する世帯像を把握することが可能ならば，その世帯に適した商品やサービスのレコメンドが可能になると考えられる．そして，世帯グループごとの活動を測定するマーケティングにも利用可能になると考えられる．

従来，ソーシャルメディアの利用履歴等から，そのユーザの性別や年齢，趣味・嗜好等のユーザ属性を推定する研究が実施されてきた [1], [2]．また，ユーザが訪問した場所の履歴からユーザ属性を推定する研究も実施されてきた [3]．しかしながら，これらの研究では，個人の行動から個人の属性を把握することを目的とし，世帯のような集団の属性は推定対象とされていない．また，移動履歴や web 閲覧履歴等，ユーザにとって提示することの心理的障壁が高い情報からユーザ属性を推定している．

本研究では，レコメンドやマーケティングでの活用が可能な，世帯像を推定するシステムの実現を目指す．世帯を表現する情報の中でまずは，世帯の基本情報である世帯人数の推定を目指す．本システムでは，世帯人数の推定に向け，ユーザが提供することに心理的障壁が高くなく，各世帯の特徴を表現可能と考えられる，機器の利用状況等の機器に関する情報を活用する．

厚生労働省調査 [4] では世帯構成を，単独世帯と夫婦のみの世帯，夫婦と未婚の子のみの世帯，ひとり親と未婚の子のみの世帯，三世代世帯と定義している．これら世帯構成は，**表 1** のように世帯人数と対応付けることが可能である．たとえば，子どもがいる世帯にレコメンドしたい場合は，3 人以上世帯にレコメンドすることで 93.2%の適合率でレコメンド可能となる（表 1）．また，可処分所得が多い単独世帯に対してレコメンドする場合は，1 人世帯に対してレコメンドすればよい．世帯人数を把握可能となれば，その世帯における消耗品の利用サイクル等も推測でき，レコメンド頻度を調整できる．また，同じ商品でも大きさや容量等を考慮した適切な情報を提供することも可能となる．したがって，人数を推定することで，レコメンドに向けた基本情報を推定することが可能となる．そして，推測した世帯の活動を測定するようなマーケティングを実施することで，その世帯グループのさらなる生活行動の理解にも応用することが可能となる．

世帯人数情報の取得に関しては，サービス契約時に，ユーザから直接世帯人数を聞く運用方法が考えられる．しかしながら，これらの調査に対し，多くの人が回答しないことが経験より分かっている．そこで，本システムでは，少数ではあるが回答のあったユーザの世帯人数情報を学習し，

表 1　世帯人数と世帯構成 [4] の依存関係
Table 1　The deep interdependence between a household size and a household composition.

世帯人数	世帯構成	割合（%）
1 人	単独世帯	26.5
2 人	夫婦のみ世帯	23.2
	ひとり親と子のみ世帯	5.2
	その他世帯	3.9
3 人以上	夫婦と子のみ世帯	29.7
	ひとり親と子のみ世帯	2.0
	三世代世帯	6.6
	その他世帯	2.8
計		100.0

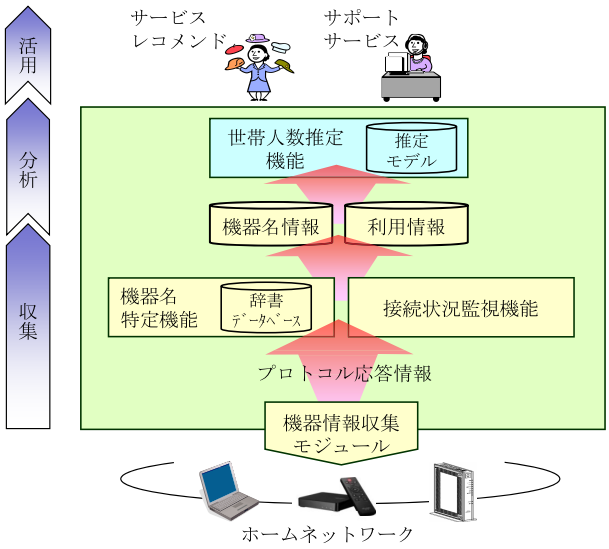

図 1　機器情報から世帯人数を推定するイメージ図
Fig. 1　The outline of the estimation flow for family information using device information.

回答のなかった大多数のユーザの世帯人数情報を推定する運用を想定する．

我々は今まで，ホームネットワークに接続された機器の機器名を特定する機器名特定技術の研究開発を進めてきた [5], [6]．本システムでは，世帯人数推定に必要なホームネットワーク内の機器に関する情報の収集において，この機器名特定技術を活用することとした．そして，機器名特定に向けて機器情報を収集するモジュール（**図 1**，機器情報収集モジュール）を，組み込み先機器の電源オンオフ状態に依存せずに，ホームネットワーク内各機器から定期的に情報を取得できるよう，電源が切れることのないサービスゲートウェイ（以下，SGW）に組み込む方針とした．機器に関する情報としては，機器名情報と利用情報を利用する．利用情報は，データ量が膨大で，各世帯における利用パターンがすべて異なるように表現され，同じ世帯人数の世帯間で類似する特徴を表現しにくくなるという課題が存在する．そこで本システムでは，世帯の特徴を表現できる

情報となるように利用情報を丸める加工を実施する．

本研究で実ホームネットワークに対して，一部である機器情報収集モジュールが組み込まれた SGW を接続し，収集サーバにてホームネットワーク内各機器から機器名情報と利用情報を取得できることを確認した．さらに，SGW を接続した複数のホームネットワークから収集した機器名情報と利用情報を学習して，世帯人数を推定する処理が問題なく実施できていることを確認した．世帯人数推定の精度評価に向け，1,000 世帯分の機器に関する情報と世帯人数に対して機械学習を実施して推定モデルを構築し，推定モデルの精度を測定した．その結果，推定対象の世帯人数の世帯のデータか否かを判定する分類において 83.7%の適合率で推定できた．その中で，商品やレコメンドに適した世帯である，可処分所得が多い 1 人世帯においては 89.5%の適合率で推定できることを確認した．また，子どもを含む世帯が多い 3 人以上世帯においては 88.3%の適合率で推定できることを確認した．本実験結果から，人数推定における，機器に関する情報の有効性を確認することができた．

本論文では，世帯人数推定を実現するシステムと，機械学習に利用した機器に関する情報の加工手法，世帯人数推定の精度について述べる．本論文の構成は，以下のとおりである．まず 2 章で，世帯人数を推定するシステムの全体構成について述べ，人数推定に利用する機器の情報と，世帯の特徴の表現に向け，利用情報を加工する処理について述べる．3 章で世帯人数推定システムにおける人数推定の精度評価結果について述べる．そして，推定に利用した各情報が精度に与えた影響についての考察を述べる．4 章で将来課題について述べ，5 章でまとめを述べる．

2. 世帯人数推定システム

本システムでは，多数のユーザ宅の所有機器に関する情報と世帯人数の組合せを機械学習することで，ある世帯の所有機器情報から，その機器を所有する世帯人数を推定する．機械学習で推定する情報（以下，目的変数）は世帯人数であり，その世帯人数を推定するために入力する情報（以下，説明変数）として，ホームネットワークに接続された機器の機器名情報と，各機器の利用情報を用いる．

2.1 説明変数獲得に向けた機器名特定技術の活用

本システムでは，説明変数として利用する機器名情報と利用情報を収集する必要がある．筆者らは，IT 資産管理を目的として，ホームネットワークに接続された機器の機器名情報と接続状態を収集可能な機器名特定技術 [5], [6] を研究開発してきた．世帯人数推定システムの実現に向け，ホームネットワークから機器に関する情報を収集可能な，機器名特定技術を活用する方針とした．

機器名特定技術では，まず，ホームネットワーク内の各機器に HTTP や UPnP，SNMP 等複数種のプロトコルの信号を送信する．そして，機器からの各プロトコルに対する応答情報を，辞書データベース（図 1．以下，辞書 DB）と照合することにより，機器名を特定する．辞書 DB とは，応答するプロトコル信号と，それに対する応答情報を機器ごとに事前に調査し，機器名情報とその応答情報のペアを保存している DB のことである．機器名特定技術で特定できる機器名情報は，以下である．

- 機器区分名
- メーカ名
- 機種名
- 型番名

本機器名情報を機械学習の説明変数として利用する．

機器名特定技術は，接続状況の監視も可能であり，発見した機器の MAC アドレスに対して定期的に ARP パケットを送信し，機器からの応答の有無を確認することで接続と切断の状況を確認する．本システムでは，この接続応答情報を機械学習の説明変数である利用情報として採用し，世帯人数の推定に活用する．

ユーザによる機器の利用情報を収集するためには，利用情報を 24 時間中定期的に取得し続ける必要がある．したがって，録画予約待機のために常時電源が入っている映像受信機 SGW からホームネットワーク内の各機器へプロトコルの信号を送信するよう設計した．本システムでは，SGW のプラットフォームとなっている android 上で各プロトコル信号を送信する，機器情報収集モジュール（**図 2**）を開発した．応答情報を辞書 DB と照合する機器名特定処理は共通化できるため，サーバで実施する設計とした．機器名特定処理を行った後の情報もサーバに蓄積する．この設計にともない，各プロトコル信号送信部から取得した情

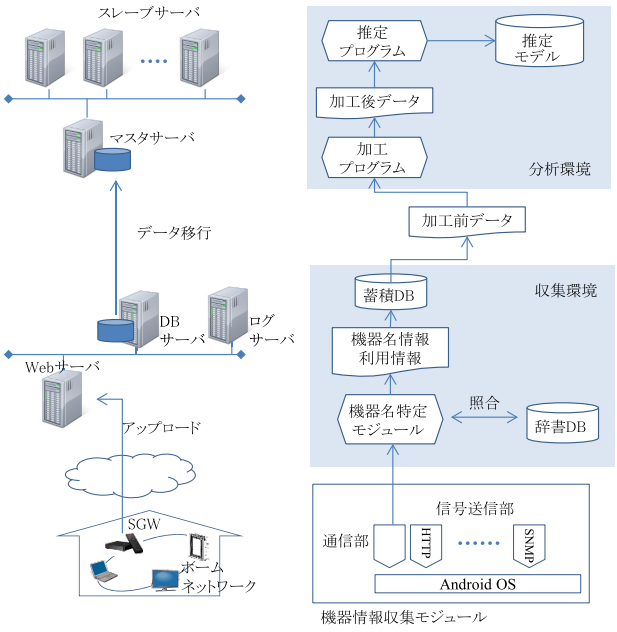

図 2　物理構成と処理の流れ

Fig. 2　System component and the process flow.

表 2　アルゴリズムパラメータ一覧
Table 2　The list of parameters used in the algorithms.

アルゴリズム	パラメータ	値
SVM	kernel	rbfdot
	sigma	PSO[12]（※）の利用による自動設定
	type	C-svc
	cross	5
Random Forest	ntree	500
	mtry	変量総数
Neural network	size	3
	decay	0.1
	weights	1
	range	0.7
	MaxNWts	4000
Logistic 回帰	family	BINOMIAL, LOGIT

※）PSO の利用パラメータは以下のとおり．
粒子数：40，試行回数：10，
w：$1/(2\log 2)$，c_1：$0.5+\log 2$，c_2：$0.5+\log 2$

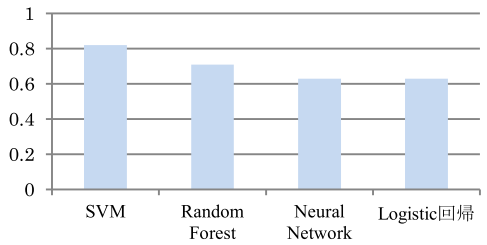

図 3　各分析アルゴリズムにおける適合率
Fig. 3　The precision of each algorithm.

報をサーバにアップロードする通信部（図 2）も開発し，SGW に組み込んだ．そして，本 SGW の初期起動時にはデータ取得に関する同意規約文が表示され，同意が得られた場合のみデータを取得する設計とした．

変数量が多い利用情報の設計前に，変数量が少ない所有機器名情報のみの場合においてもより精度が高い機械学習アルゴリズムを選定するため，主要な機械学習アルゴリズムに対して簡易な比較実験を実施した．世帯人数と機器区分ごとの合計台数，メーカごとの台数についてサンプルデータを取得し，3 人世帯の推定（二値分類）を各アルゴリズムで行った．比較を行ったアルゴリズムは SVM [8], Random Forest [9], Neural Network [10], Logistic 回帰 [11] であり，それぞれのパラメータは**表 2** に示すとおりである．実験の結果は**図 3** に示すとおり SVM の適合率が最も高かったため，本システムではまずは SVM を利用することとした．また，SVM のアルゴリズムパラメータを最適化する PSO 手法 [12] もあわせて適用することとした．利用する学習アルゴリズムについては，今回実験により判定したが，今後説明変数の見直し時等に各アルゴリズムの数理モデルと比較して，最適なアルゴリズムに順次切り替えていく予定である．

2.2　説明変数加工

世帯の特徴を表す利用情報のデータは，定義した機器区分数や，メーカ数と学習時間数を乗じた数だけ存在することになり，利用情報のデータ量は膨大になってしまう．各世帯における利用情報のパターンがすべて異なるように表現されることで，同一世帯人数内の世帯間でもデータの類似性が低下することが考えられ，機械学習のアルゴリズムによっては集合としての特徴を抽出できず，推定モデルの精度が低下する懸念が高くなる．

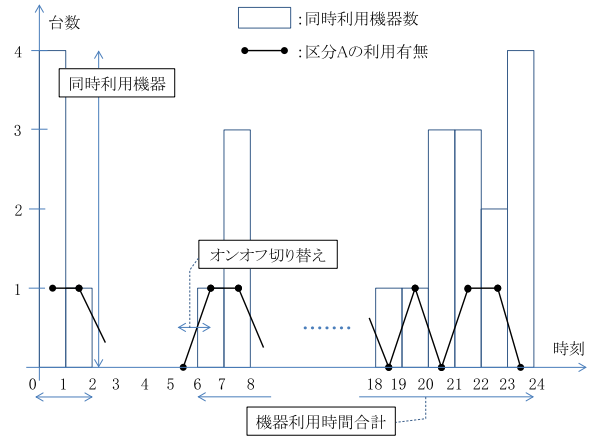

図 4　抽出する利用情報
Fig. 4　The usage status information.

そこで，本システムでは，同一世帯人数の各世帯間で類似する特徴を表現可能とするため，以下のように利用情報を丸める加工を実施する．

(1)　最大同時利用機器数

世帯人数が同じ場合，同時に利用している機器の合計台数も似ていると考えられる．そこで，1 日における同時利用機器数（**図 4**）の最大値を説明変数とする．これは，単位時間ごとに同時利用している機器数の総和を求め，さらにそれらの中で 1 日ごとの最大値を計算することで求める．

図 4 のように単位時間を 1 時間とすると，本変数は以下の式で表すことができる．

$$\max \left[\begin{array}{c} \text{0:00〜1:00 までの利用機器合計数,} \\ \text{1:00〜2:00 までの利用機器合計数,} \\ \cdots \\ \text{23:00〜24:00 までの利用機器合計数} \end{array} \right]$$

利用機器合計は，ホームネットワーク内の各機器に対して定期的に送信する ARP に対しての応答数の単位時間内の最大値である．

(2)　時間帯ごとの機器利用時間合計

表 1 のように世帯人数は世帯構成と関連付けることができる．3 人以上世帯は子供がいる世帯が多いため，午後や夕方から機器の利用が始まると想定される．一方，2 人世帯で最も多い世帯である，夫婦のみ世帯では日中 2 人とも仕事で家にいないことが想定され，夜から機器が利用され始めると想定される．また，1 人世帯の場合は深夜時間帯においても機器が利用され続けると想定される．したがっ

て，世帯人数が同じ場合，機器の利用時間帯も似ていると考えられる．そこで，各時間帯において機器を利用していた時間（図 4）の合計値を説明変数とする．これは，単位時間ごとに所有有無の論理和をとり，6 時間のブロックで区切った時間内での和を計算することで求める．文献 [7] によると機器の利用のピークが休日・平日ともに朝 7 時前後，夜 22 時前後であるため，この特徴を抽出できるよう，6:00～12:00 の午前時間帯，18:00～24:00 の夕・夜時間帯と，それ以外の 12:00～18:00 の午後時間帯，24:00～6:00 の深夜・朝時間帯の 6 時間ごとのブロックを利用した．

単位時間を 1 時間とすると，6:00～12:00 時間帯の機器利用時間合計は，

(6:00～7:00 の機器利用) + (7:00～8:00 の機器利用)
+ · · · + (11:00～12:00 の機器利用)

となる．機器利用とは，単位時間内におけるホームネットワーク内全機器の利用有無であり，ホームネットワーク内機器数が n のときは，以下の式で表すことができる．

(· · · ((機器 1 の利用有無 or 機器 2 の利用有無) or
機器 3 の利用有無) · · · or 機器 n の利用有無) · · ·)

利用有無とは，単位時間のうち，一度でもその機器の利用を確認すると 1 となる．一方，単位時間のうち，一度もその機器の利用を確認できない場合は 0 となる．

(3)　オンオフ切替え回数合計

世帯人数が同じ場合，機器の利用頻度も似ていると考えられる．そこで，機器区分ごとの利用状況の切り替わり回数を説明変数とする．これは，単位時間ごと，機器区分ごとに論理和をとり，1 日単位で「0」から「1」へ，「1」から「0」へ切り替わった回数を計算することで求める．

単位時間を 1 時間として，A という機器区分のオンオフ切替え回数合計は以下の式で求められる．

|(0:00～1:00 の区分 A の利用有無)
− (1:00～2:00 の区分 A の利用有無)|
+ |(1:00～2:00 の区分 A の利用有無)
− (2:00～3:00 の区分 A の利用有無)| + · · ·
+ |(22:00～23:00 の区分 A の利用有無)
− (23:00～24:00 の区分 A の利用有無)|

区分 A の利用有無については，単位時間内で一度でも区分 A のいずれの機器の利用を確認した場合 1 となり，一方，区分 A のいずれかの機器の利用を一度も確認できなかった場合 0 となる．

上記を利用情報として活用し，機械学習を実施する．

3.　世帯人数推定システムにおける推定の精度評価

本研究では，SGW を所有しているユーザにおいて，SGW からホームネットワーク内機器に関する情報を収集してサーバに送信し，サーバで収集した情報を学習して推定モデルを構築するシステムを開発した．実ホームネットワークにおいて，機器情報収集モジュールが組み込まれた SGW を接続し，ホームネットワーク内各機器から機器名情報と利用情報を取得できることを収集サーバ群で確認した．さらに，SGW を接続した複数のホームネットワークから収集した機器名情報と利用情報を学習して，世帯人数を推定する処理が問題なく実施できていることも確認した．

本システムの商用利用前に，アンケートデータをもとに本システムの世帯人数推定の精度を評価した結果について述べる．

3.1　構築する推定モデルと評価内容

ある特定の目的変数とそれ以外の目的変数のどちらに適合するかを判定する二値分類の推定モデルを構築する．この推定モデルは，レコメンドやマーケティングのように，推定対象がターゲットとなる対象であるか否かを判定する．そして，世帯の人数を他サービスへ流通させるビジネスのように，ある世帯の人数を一意に推定する多値分類の推定モデルも構築する．

本研究では，マーケティングのように特定の世帯の活動を測定するユースケースも見据えている．マーケティングでは，対象となる世帯に限定して活動を調査する必要があるため，推定結果の誤りが少ない方が良い．したがって，本評価では適合率を重視して適合率を評価する．

3.2　利用データ

本システムにおける人数推定の精度を評価するため，1,000 世帯に対して web アンケートを実施し，世帯人数と世帯構成，所有機器の機器名情報，利用情報の正解データを取得した．世帯人数の比率を国民生活基礎調査 [4] の結果に合わせ，1 人世帯から 6 人世帯までを選出して web アンケートを実施した．7 人以上世帯は，全体の 4.8% [13] と少数であったため，本調査では除外した．所有機器を実際にホームネットワークに接続しているか否かは問わず，ネットワークに接続可能な機器区分をあらかじめアンケートに用意しておき，その機器区分に合致する機器のみ，メーカ名と機種名，型番の情報を入力してもらった．白物家電の多くはネットワーク接続機能がないため，白物家電についての所有情報はアンケートで取得せず，多くの機器がネットワーク接続機能を有する黒物家電とタブレット機器等について所有情報を取得した．アンケートにて取得した所有機器の機器区分は以下の 7 つである．

① 携帯電話（スマートフォン含む）
② パソコン
③ TV
④ ゲーム機

表 3　説明変数の例
Table 3　The example of explanatory variables.

	機器名情報			利用情報													
	全台数	台数	所有有無	最大同時利用機器数				機器利用時間合計								オンオフ切り替え回数合計	
								平日				休日				平日	休日
				1 日目	2 日目	3 日目	4 日目	午前	午後	夕夜	深夜	午前	午後	夕夜	深夜		
全機器	8			4	4	6	7										
①携帯電話		3	1					4	4	6	1	4	5	6	2	6	4
A1 社製		0	0														
B1 社製		2	1														
··																	
②パソコン		1	1					1	0	2	0	0	1	2	1	4	4
A2 社製		0	0														
B2 社製		1	1														
··																	
··																	
⑦ネットワーク機器		1	1					1	1	1	1	1	1	1	1	0	0
A7 社製		1	1														
B7 社製		0	0														
··																	

⑤ タブレット
⑥ プリンタ
⑦ ネットワーク機器
（無線 LAN アクセスポイント，スイッチングハブ等）

上記の機器の製造メーカとして 154 社のメーカ名データを用意した．利用情報については，木曜日から日曜日まで 4 日間の連続 96 時間分の利用状況を，1 時間単位で機器ごとに取得した．1 時間のうち 1 回でも使用した場合，使用した旨を申告してもらい，1 時間のうち 1 回も使用しなかった場合は，使用しなかった旨を申告してもらった．

所有機器に関する本アンケートでは，メーカ名に関してはプルダウン形式で用意したが，機種名と型番は自由記述とした．その結果，機種名と型番については空白であることが多く，回答率が悪かった．したがって，本評価では機器区分名とメーカ名の情報のみを機器名情報として利用することとした．機器名情報としては全機器台数と，機器区分ごとの台数と所有有無，機器区分内のメーカごとの台数と所有有無を変数として用意した．具体的な説明変数の例を**表 3** に示す．所有有無情報は，所有台数は一致しなくても，所有状況が一致することを特徴化するため用いた変数であり，1 台でも所有していれば数字の 1 を値として持ち，1 台も存在しない場合は数字の 0 を値として持つ（表 3）．表 3 では，あるメーカの機器のみを行に記載しているが，実際のテーブルではすべて（154 社）の機器メーカの行を持っており，所有していない機器については台数・所有有無ともに「0」となる．

3.3　利用情報に対する丸め処理の実施

本評価では，木曜日から日曜日までの 4 日間のデータを取得した．2.2 節で述べた機器利用情報の丸め加工処理の結果，(1) 最大同時利用機器数はデータが 4 日分あるため，説明変数の数は 4 日分で 4 となった．(2) 時間帯ごとの機器利用時間合計は，木曜日と金曜日のデータを平日のデータ，土曜日と日曜日のデータを休日のデータとして，機器区分ごとに平日と休日の 6 時間ごと各々 4 つの時間帯について使用時間合計を取得し，2 日分の平均をとった．説明変数の数は 7 機器区分，8 時間帯分の組合せで 56 となる．休日と平日のデータを分離した理由としては，同じユーザであったとしても休日と平日で機器の利用形態が異なるためである [7]．(3) オンオフ切替え回数合計は，機器区分ごとに，平日と休日につきオンオフ切替え回数を取得し，説明変数の数は 7 機器区分の 2 日分の 14 となった．

3.4　適合率算出方法

1,000 世帯分のデータをランダムに 5 グループに分類し，そのうち 4 グループを推定モデル構築用途で使用し，残りの 1 グループを推定モデル評価用データとして使用するクロスバリデーション方式 [14] を採用した．具体的には，1,000 世帯を 200 世帯ずつにランダムに分けて，4 グループ分の 800 世帯を対象に学習を行い，推定モデルを構築する．その後，残りの 200 世帯に対してその推定モデルを適用し，200 世帯中に含まれる目的変数の世帯を集計して，適合率を求める．次に評価用データとして利用したグループを推定モデル構築用に使用し，推定モデル構築用に使用したグループの中から評価用として検証するグループを 1 つ選択する．全グループにおいて評価用データとして，再実験していく．グループ数である，全 5 回の評価の適合率の平均を，最終的な適合率として算出する．

表 4　各目的変数における適合率
Table 4　Precision for each objective variable.

目的変数	適合率（%）		正解データ数
	二値分類	多値分類	
1 人	89.5	63.6	176
2 人	85.7	60.4	255
3 人	78.8	61.4	275
4 人	93.3	64.7	209
5 人	54.5	85.7	57
6 人	66.7	100.0	28
加重平均（※）	83.7	64.4	1,000（合計）

※）加重平均とは，値の重みを加味して平均すること．
この場合，重みは正解データ数と言い換えることができる．

3.5　世帯人数推定結果

クロスバリデーションの 5 パターンのすべてにおいて，各推定モデルを構築することができた．世帯人数推定の結果は，**表 4** に示すとおりである．ある特定の目的変数とそれ以外の目的変数のどちらに適合するかを判定する二値分類の適合率の加重平均が 83.7%であった．さらに，商品やサービスのレコメンドに適した世帯である，可処分所得が多い 1 人世帯においては 89.5%の適合率で推定できることを確認した．世帯の人数を一意に推定する多値分類においては，適合率の加重平均が 64.4%であった．二値分類における 5 人，6 人の推定結果は，適合率向上のために，正解データ数が少ない不均衡なデータを調整（SMOTE [15] を活用）した結果である．一方，多値分類時においても 5 人，6 人の推定時に同様に不均衡データを調整した場合，適合率が低下する事象が発生した．二値分類器での不均衡データの調整が，多値分類において過学習の傾向を強めてしまったと考えられる．したがって，多値分類への利用時には不均衡データの調整は実施しなかった．この処理によって，多値分類の結果が二値分類よりも上回ったと考えられる．

子ども世帯が多い，3 人以上世帯の判定においては，多値分類において 3 人〜6 人と推定できれば正解と判定すると推定は 88.3%で特定できた．高い精度で対象とする世帯人数か否かを判定することができ，世帯人数の推定において，機器に関する情報の有効性を確認することができた．

3.6　世帯人数推定に関する考察

3.6.1　利用情報に対する丸め処理の効果

利用情報に対して，丸め処理を行った場合と，丸め処理を行わなかった場合で適合率を比較することで，丸め処理の効果を確認した．子ども世帯が多い 3 人以上の判定を，多値分類で 3 人〜6 人と判定されたか否かで世帯判定できる等，世帯像の類推が可能で，応用領域が広いため多値分類にて比較を行った．適合率の加重平均は，丸め処理を行わなかった場合 54.4%であり，丸め処理を行った場合 58.1%で 1 割程度の向上を確認することができた．この結果より，丸め処理の効果を確認することができた．なお，本実験は説明変数の組合せと適合率の変化の関係性を一次評価するために行ったものであり，実験時間を要する PSO は適用しないこととした．

表 5　各説明変数の組合せと適合率の平均
Table 5　The variation of explanatory variables and the average of precision.

説明変数の組み合わせ	6 目的変数の適合率（%）の加重平均
ホームネットワーク内機器の全台数のみ(ⅰ)	44.7
機器区分ごとの台数と所有有無のみ(ⅱ)	62.3
機器区分内のメーカごとの台数と所有有無のみ(ⅲ)	66.5
(ⅰ)+(ⅱ)+(ⅲ)+利用情報（3.5 節の結果と同じ）	83.7

3.6.2　機器名情報の詳細度と適合率の関係

今後の説明変数，すなわち，ユーザから収集する情報に関する検討に向けて，機器名情報の詳細化の効果について考察する．下記 (i) から (iii) の項目になるに従い，機器に関しての情報が詳細になっている．

(i)　ホームネットワーク内機器の全台数
(ii)　機器区分ごとの台数と所有有無
(iii) 機器区分内のメーカごとの台数と所有有無

この機器の詳細度が適合率に与える影響に関する知見の獲得を目的に，それぞれの情報のみを用いた際の適合率を評価する．

(i) ホームネットワーク内機器の全台数，(ii) 機器区分ごとの台数と所有有無，(iii) 機器区分内のメーカごとの台数と所有有無のそれぞれの説明変数で各目的変数に対して二値分類推定を実施した際の適合率の加重平均は**表 5** のようになった．(i) ホームネットワーク内機器の全台数から (iii) 機器区分内のメーカごとの台数と所有有無のみへ，詳細化が進むに従い適合率が向上している．

特に，(i) ホームネットワーク内機器の全台数と (ii) 機器区分ごとの台数と所有有無の適合率の差が大きかった．これは，世帯人数が同じであれば，スマートフォンやタブレット等，複数台持つ機器区分が同じであり，かつその機器区分における台数も同じであったためだと推測できる．さらに，(ii) 機器区分ごとの台数と所有有無と，(iii) メーカごとの台数と所有有無でも少し差が出た理由については，世帯人数が同じ場合，選ぶ家電の容量や大きさが類似するため，所有するメーカが同じであったためと推測できる．この結果より，機器名情報については，詳細度を高めることで適合率が向上することを確認できた．また，上記 (iii)

の説明変数に加え，丸め加工処理を行った利用情報を加えた際も，適合率の結果が向上した．この結果より，丸め処理を行った利用情報の有効性を確認することができた．

4. 将来課題

4.1 目的変数の多様化

本評価においては目的変数として，世帯人数を設定したが，世帯属性としては他にも非常に多くの属性項目が想定される．たとえば，世帯属性情報として，世帯の各構成員の年齢，性別がある．世帯の各構成員の年齢や性別を表現することにより，さらに詳細に世帯像を把握することが可能になると考えられる．

年齢や性別は，所有する機器区分や機種，型番に特徴が出ると推測できる．若年層にはデジタル家電と呼ばれるスマートフォンやタブレットが受け入れられ，かつ，ホームネットワークの利用においても無線 LAN アクセスポイントを利用する傾向にある [16]．一方，高齢層にはそのような傾向は見られない．また，同じ機器区分においても若年層にはデザイン性の高い機種が受け入れられ，高齢層には使いやすい機種が受け入れられやすい傾向にある [17]．性別においては，好みの色で特徴付けられると想像できる．これら色情報については多くの場合で機器の型番に情報が含まれている．

したがって，機器に関する情報として機種名や型番まで利用して表現することで，世帯像の特徴をより表現できるようになると期待できる．3.6 節で示したように機器名情報の詳細度を高めることで適合率が向上するという検証結果が得られたため，今後は機器区分名とメーカ名に関する情報だけではなく，機種名や型番のレイヤまで情報を細分化して，機器名の詳細度を高めることで，さらなる世帯像の把握を目指していきたい．

4.2 利用情報の精度向上

今回の評価では，機器の利用状況をアンケートにより取得したが，実際に機器の利用状況を取得する際は，SGW から ARP 応答要求信号を機器に送信し，その応答信号が返ってきた場合，その機器は利用されていると判断し，応答信号が返ってこない場合は利用されていないと判断することを想定している．昨今の機器では，利用していないときに自動で省電力モード（ただし，電源はオン状態のまま）に移行し，利用されていない状態でも，ARP の応答要求信号に応答するため，正しく利用状況を把握することが不可能な機器も多く存在する [18]．今後，正しくユーザの利用状況を取得するため，電源がオン状態であったとしても，利用している状態と利用していない状態を分離していく必要がある．

5. まとめ

本論文では，ホームネットワークに接続された機器の機器名情報と利用情報から，機器を所有している世帯の人数を推定するシステムについて述べた．本システムでは，膨大な利用情報に対して，世帯の特徴を表現できるよう丸め処理を実施し，多数の世帯からの情報を学習して推定モデルを作成することで，世帯人数を推定する．本システムを実ホームネットワークに適用し，各ホームネットワークから機器名情報と利用情報を収集し，世帯人数推定のモデルを構築できることを確認した．本システムによる世帯人数推定の評価として，1,000 世帯から機器名情報と利用情報，世帯人数情報をアンケートにより収集し，世帯人数の推定を実施した．その結果，ある特定の目的変数とそれ以外の目的変数のどちらに適合するかを判定する二値分類にて平均 83.7%の精度で推定できた．そして，商品やサービスのレコメンドに適した世帯である，可処分所得が多い 1 人世帯においては 89.5%の適合率で推定でき，子どもを含む世帯が多い 3 人以上世帯においては 88.3%の適合率で推定できた．機器に関する情報から，高い精度で世帯の人数を推定できたため，世帯人数推定における機器に関する情報の有効性を確認することができた．また，本評価において，丸め処理を行った利用情報を説明変数に利用しなかった場合と比較して，適合率の向上を確認できたため，世帯人数推定において丸め処理した情報の有効性も示すことができた．

今後は，各機器の利用情報を正しく把握する手法等，将来課題に対応していくだけでなく，情報を取得する際のユーザへのパーミッションのあり方についても検討してきたい．

参考文献

[1] 蔵内雄貴，内山俊郎，内山　匡：マルコフ確率場を用いたソーシャルネットワークからのユーザ属性推定，電子情報通信学会論文誌 D，情報・システム，J96-D(6), pp.1503–1512 (2013).

[2] 伊藤　淳，西田京介，星出高秀：Twitter と Blog の共通ユーザおよび会話ユーザの同類性に着目した Twitter ユーザ属性推定，日本データベース学会論文誌，Vol.12, No.1, pp.31–36 (2013).

[3] 篠田裕之，竹内　亨，寺西裕一，春本　要，下條真司：行動履歴に基づく協調フィルタリングによる行動ナビゲーション手法，情報処理学会研究報告，GN2007(91), pp.87–92 (2007).

[4] 厚生労働省大臣官房統計情報部：グラフで見る世帯の状況（国民生活基礎調査（平成 22 年）の結果から）(2012).

[5] 美原義行，山本隆二，佐久間聡，山崎毅文，岡本　学，佐藤敦：ユーザ端末を対象とした機器名特定システムの開発，情報処理学会論文誌コンシューマ・デバイス&システム（CDS），Vol.3, No.1, pp.64–76 (2013).

[6] 浅野貴久，美原義行，高谷太紹，小林昭久，山口徹也，高倉健：ビッグデータ利活用における課題と今後のサービス活用に向けて，電気通信，Vol.76, No.799, pp.25–34 (2013).

[7] 総務省：情報通信白書平成 25 年度版 (2014).
[8] Vapnik, V. and Cortes, C.: Support vector networks, *Machine Learning*, Vol.20, pp.273–297 (1995).
[9] Breiman, L.: Random forest, *Machine Learning*, Vol.45, pp.5–32 (2001).
[10] Rosenblatt, F.: The Perceptron: A Probabilistic Model for Information Storage and Organization in the Brain, *Psychological Review*, Vol.65, No.6, pp.386–408 (1958).
[11] Dawber, T.: *The Framingham Study: The Epidemiology of Atherosclerotic Disease*, Cambridge, Mass: Harvard University Press (1980).
[12] Clerc, M.: Standard Particle Swarm Optimisation (2011), available from ⟨http://clerc.maurice.free.fr/pso/SPSO-descriptions.pdf⟩ (accessed 2015-12-09).
[13] 総務省統計局：日本統計年鑑第二章人口・世帯．Seymour, G.: *Predictive Inference*, CRC Press, ISBN 0-412-03471-9 (1993).
[14] Seymour, G.: *Predictive Inference*, CRC Press, ISBN 0-412-03471-9 (1993).
[15] Nitesh, V., Chawla, 1., Kevin, W., Bowyer, Hall, 1 L.O. and Kegelmeyer, W.P.: SMOTE: Synthetic Minority Over sampling Technique, *Journal of Artificial Intelligence Research*, Vol.16, pp.321–357 (2002).
[16] 博報堂生活総研オンライン：若者の消費特性と価値観—買わない若者の背景を考える (May 2015)，入手先 ⟨http://seikatsusoken.jp/teiten/theme/report-35/⟩(参照 2015-12-09)．
[17] 日本電機工業会家電調査委員会：団塊世代の白物家電に対する調査 (2008).
[18] 高谷太紹，美原義行，小林昭久，山口徹也：ホームネットワーク接続機器の状態把握に関する提案，第 76 回全国大会講演論文集 (2014).

美原 義行　（正会員）

NTT サービスエボリューション研究所．2004 年東京工業大学理学部情報科学科卒業，2006 年同大学院情報理工学系研究科数理・計算科学専攻修了．2006 年 NTT 入社．以来，ホームネットワーク管理サービスの技術設計等の研究開発，プロトコルの標準化に従事．現在，NTT サービスエボリューション研究所研究主任．

山口 徹也

NTT サービスエボリューション研究所．1997 年大阪大学工学部情報システム工学科卒業，1999 年同大学院工学研究科電子情報エネルギー工学専攻博士前期課程修了．同年 NTT 入社．以来，コンテンツナビゲーション技術，IPTV 伝送技術，ホームネットワーク管理システム，超高臨場メディア伝送技術の研究開発に従事．現在，NTT サービスエボリューション研究所主任研究員．博士（情報科学）．電子情報通信学会会員．

高倉 健

NTT サービスイノベーション総合研究所．1990 年大阪大学基礎工学部卒業．1992 年同大学院基礎工学研究科物理系専攻修士課程修了．1992 年 NTT 入社．以来，映像情報サーバ，情報アクセス制御方式および分散データ管理システムの研究開発に従事．現在，NTT サービスイノベーション総合研究所研究推進担当部長．