Vol.2014-MBL-72 No.6
Vol.2014-CDS-11 No.2
2014/8/28

ホームネットワーク内接続機器の情報を活用した世帯属性推定手法

美原義行† 山口徹也† 高倉健†

本研究では，ホームネットワークに接続された機器から取得可能な情報を用いて，世帯人数や世帯構成等の世帯属性を推定する手法を提案する．そして，提案手法の適合精度を評価することで，世帯属性の推定に対する機器から取得可能な情報の有効性を検証する．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であり，その情報を機械学習し特徴を把握することで，世帯属性を推定する．世帯属性を把握することにより，その世帯に適した機器の連携サービスの提供が期待できる．評価において，1,000 世帯に対するアンケートから機器名情報と利用情報，世帯属性情報を収集・学習し，世帯属性の推定を実施した．その結果，二値分類で平均 83.7%，多値分類で平均 64.4%の制度で推定でき，ランダムに推測するよりも精度高く推定でき，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，本評価において機械学習の説明変数の種類を多くするほど，かつ，粒度を細分化するほどに適合率が向上することを確認でき，学習時に利用する情報の設計に有効な知見を獲得することができた．

A Method Which Estimates Family Attributes using Information from Devices Connected to Home Network

YOSHIYUKI MIHARA† TETSUYA YAMAGUCHI† TAKESHI TAKAKURA†

In this paper, we propose a method which estimates the family attributes using information from devices connected to home network. Moreover, we verify practical effectiveness of the information for the estimation of the family attributes by testing prediction rates. Information which we can get from devices connected to home network is device name, such as, device type or manufacturer name, and usage status information of each device. Our method estimates the family attributes by learning data from devices and by detecting features of each data. Estimating the family attributes allows us to provide service appropriate to the family. We carry out a questionnaire on 1,000 family attributes, device name and the usage status information. We can show the effectiveness of the device information with four times precision by comparing random estimation. Moreover, we can accumulate knowledge which precision rates improve with explanatory variables becoming diversified and reducing level of abstraction.

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，機器同士を連携させるサービスも拡充されつつある． さらに，ホームネットワークの所有する世帯の人数や構成といった，世帯属性を把握することが可能ならば，その世帯にとって，より魅力的なサービスの提供が可能になると考えられる．例えば，共働き子育て世帯には，留守中の子供のために窓開閉センサ等を活用した見守り・防犯サービスが有効であると考えられる．また，高齢者世帯には，高齢者は一般的に煩雑な操作が苦手とされるため，オペレータによる遠隔からの機器操作代行サービスが有効であるだけでなく，サポートするオペレータが年齢等の属性を踏まえた応対も行える．上記のような，世帯に適したサービスを推薦したり，プロアクティブに提供するためには，その世帯の属性を把握することが有効である．

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測する研究が実施されている[1][2]．また，ユーザが移動して訪問した場所の移動履歴からユーザ属性を推測する研究も実施さ

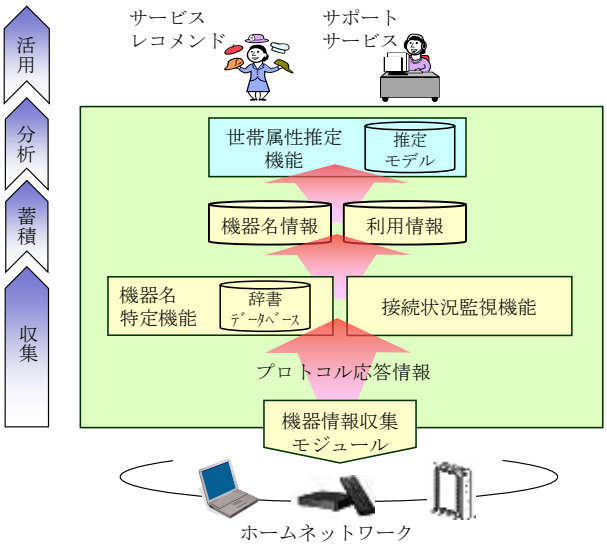

図 1 機器情報から世帯属性を推定するイメージ図
Figure 1 The outline of the estimation flow for family information using device information.

†西日本電信電話株式会社 研究開発センタ
NTT WEST CORPORATION.

れている[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測対象とされていない．本研究では，世帯人数や世帯構成等の世帯の属性を推定することを目的とし，ホームネットワークに接続された機器の種別や台数，利用状況が世帯属性と相関があるという仮説を立て，機器に関する情報から世帯属性を推定する手法を提案し，有効性を評価する．

本稿の構成は，以下の通りである．まず 2 章では，機器の情報から世帯属性を推定する際の課題と，それを解決する手法を提案する．3 章では，提案手法の評価内容について述べ, 4 章で評価結果と考察について述べる．5 章で将来課題について述べ，6 章でまとめを述べる．

2. ホームネットワーク内接続端末の利用情報を活用した世帯属性推定手法の提案

本研究では，世帯属性を推定することを目標に，多数のユーザ宅の所有機器に関する情報と世帯属性の組み合わせを機械学習することで，ある機器情報を入力として，その機器を所有する世帯の属性を推定する手法を提案する（図 1）．機械学習にて推定する情報（以下，目的変数）は世帯属性であり，その世帯属性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，機器が使用された利用情報を用いる．

2.1 説明変数の取得方法

機器名情報と利用情報は機器名特定システム[4][5]を利用することで，取得することが可能である．機器名特定システムには，ホームネットワーク内の機器の機器名を特定する機能（図 1，機器名特定機能），機器の接続状況を監視する機能（図 1，接続状況監視機能）の 2 つがある．

機器名特定機能では，機器名を特定するため，ホームネットワーク内の各機器に複数種のプロトコルの信号を送信し，その信号に対する応答情報を辞書データベースと適合させることにより，機器名を特定する．辞書データベースとは，機器ごとに，応答するプロトコル信号と，それに対する応答情報を事前に調査し，機器名情報とのペアで保存しているデータベースのことである．特定できる機器名情報としては，以下のような情報がある．

- 機器区分名
- メーカ名
- 機種名
- 型番名

接続状況監視機能では，発見した MAC アドレスに対して定期的に ARP パケットを送信し, 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．本研究では，この接続情報を，ユーザによる機器の利用情報として説明変数に適用する．

2.2 説明変数と目的変数の特徴と課題

説明変数に用いた機器名情報と利用情報，目的変数である世帯属性には，それぞれ以下の特徴がある．

A) 説明変数の数が膨大

B) 正解データ数が少ない目的変数が存在

説明変数として，「機器の全台数」に始まり，「携帯電話の全台数」，「携帯電話の有無」，「A 社携帯電話の台数」，「A 社携帯電話の有無」のように，各機器区分の機器の有無，台数，さらにはメーカ名等が説明変数として考えられる．また上記の台数情報や有無情報だけでなく，各機器の利用情報も存在するため，変数の数は膨大になる．例えば機器区分として，携帯電話，パソコン，TV，ゲーム機，タブレット，プリンタ，ネットワーク機器を想定し，メーカが計 154 種類あったとし，利用情報として 1 機器に対して，4 日分（96 時間）の利用情報を想定した際，154 種類分の台数と有無, 96 の利用情報のため, 計 29,568（=154×2×96）の説明変数が存在することになる．

これら変数の最適な組み合わせは，全変数に対して説明変数として利用するか否かを組み合わせ，その際に構築した推定モデルによる世帯属性推定の適合率を比較していくことで把握可能である．全ての組み合わせの中で，適合率が最大となる説明変数の組み合わせが，その目的変数に対して精度が良い説明変数の組み合わせと判断できる．このとき組み合わせの数は，全説明変数に対して利用するか否かを組み合わせた，2 に説明変数の数を乗じた数となる．したがって，説明変数の組み合わせの全パタンの適合率を計算していく処理は計算時間が非常に膨大になる．

また，各機器の所有情報と利用情報の組み合わせ数が膨大であるため，各世帯における各データが全て異なり，データ間の類似性がなくなることが考えられる．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，推定モデルを構築できない懸念がある．

一方，該当の目的変数の正解データ数が少ないために，機械学習アルゴリズムによっては推定モデルを構築できない場合も存在する．例えば，目的変数が世帯人数であった際，世帯人数が 7 人以上の世帯から情報を得ることは，そのような世帯が少ないため困難である．

以上のように，上記(A)と(B)の特徴から，以下の課題が発生する．

課題(A)-1：説明変数の組み合わせ試行回数の増大

課題(A)-2：各世帯間のデータの類似性の減少

課題(B)：少ない正解データによる推定モデルの構築が困難

これら課題を解決する手法を以下で述べる．

2.3 世帯属性推定手法の提案

2.3.1 説明変数の組み合わせ試行回数の増大への対応（課題(A)-1）

適合率向上に向けて，説明変数の利用／非利用の組み合

わせを求めていくことは，非常に時間を要する．したがって，本研究では，機械学習に利用する変数を選択し，その変数を増加させた際の適合率を比較し，利用する変数を選択していく変数増加法[6]を利用する．適合率が向上した際は，その変数を説明変数に含め，向上しない場合はその説明変数を加えない（図 2）．本手法では，まず，ある目的変数において「機器全台数」等，基本的な変数のみを説明変数として推定モデルを構築し，推定を実施する．次に，他変数を説明変数に加えて適合率を算出する．例えば，「携帯電話の全台数」を説明変数に加え，「機器全台数」と「携帯電話」の台数の 2 つの説明変数において適合率を算出する．「携帯電話の全台数」を加える前後で適合率が低下した場合は，その目的変数においては「携帯電話の全台数」は説明変数に加えない．この処理を全変数に対して実施し，ある目的変数における適切な説明変数となる組み合わせを求める．本手法は，全パタンにおいて推定モデルを構築し適合率を求めていく手法と比較しても適合率が大幅に低下しない[6]．本手法では，適切な説明変数を求める試行回数は最大で，$\sum_{i=1}^{\text{説明変数の数}} i$ になるため，2 に説明変数の数を乗じた数から大幅に試行回数を減少させることが可能となる．

2.3.2 各世帯間のデータの類似性の減少への対応（課題(A)-2）

課題(A)-2 に対応するため，以下のように利用情報を丸めて，変数の数を削減することで，世帯間で同じデータとなる状況を発生させ，各集合の特徴を学習した推定モデルを構築する．

(1) 最大同時利用機器数

同時に利用している機器の台数は，世帯構成との関連が高いと考えられる．したがって，時間ごとに同時利用している機器数の総和を求め，1 日ごとに最大値を取得し変数とする．つまりは，24 時間中最大になる同時利用機器数を変数とする．例えば，データが 4 日分ある際は，説明変数の数は 4 日分で 4 となる．

(2) 時間帯ごとの使用時間合計

時間帯ごとの機器の利用状況を把握することで，その世帯のライフスタイルを把握することができると考えられる．したがって，1 時間ごと，機器区分ごとに論理和を取り，6 時間で区切った時間内での和を求める．24 時間を 6 時間ごとの 4 グループに分けることで，深夜・朝時間帯，午前時間帯，午後時間帯，夕・夜時間帯の生活リズムを踏まえた推定モデルを構築できる．

(3) オンオフ切り替え回数合計

機器の利用頻度を把握することで，その世帯のライフスタイルを把握することができると考えられる．したがって，機器区分ごとの利用状況の切り替わり回数を測る．1 時間ごと，機器区分ごとに論理和を取り，1 日単位で「0」から「1」へ，「1」から「0」へ切り替わった回数を測る．

2.3.3 少ない正解データにおける推定モデルの構築が困難（課題(B)）

全取得データの中で目的変数の正解データが少なく，推定モデルを構築できない場合に，正解ダミーデータを予測的に作り出す SMOTE 関数[7]を利用した．SMOTE 関数では正解ダミーデータとして，他の正解データの平均的な値を取るようにして作成する．例えば，目的変数が「世帯人数 6 人」である場合，正解データの中から 2 つを抽出し，その 2 つの正解データの全機器台数が 6 台と 10 台である場合，その 2 つのデータの中間である，全機器台数が 8 台の正解ダミーデータを作成する．この変数以外の変数に対しても，中間的な値を計算して正解ダミーデータを構築していく．本研究では，SMOTE 関数を活用して正解データが

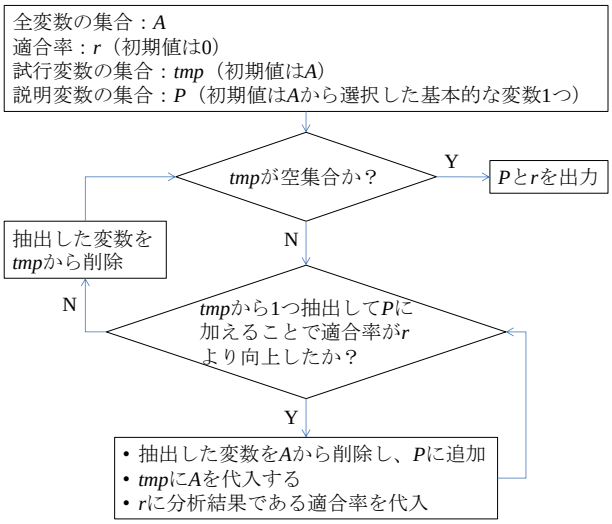

図 2 変数を増加させるフローチャート図

Figure 2 The flow chart for the step-up procedure

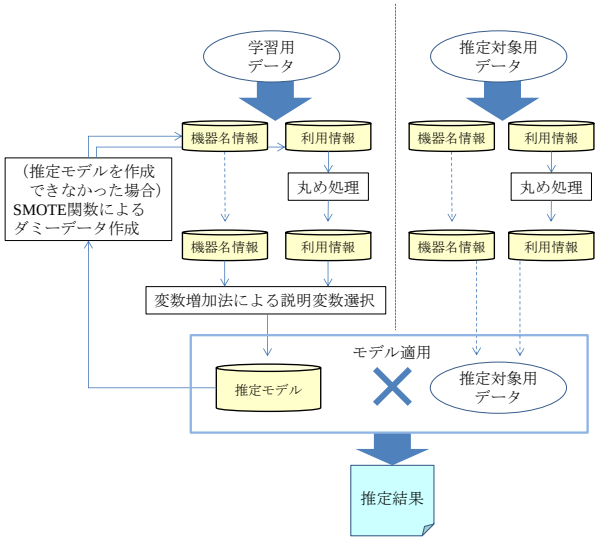

図 3 提案手法の処理の流れ

Figure 3 The flow of our proposal method

不正解データと同じ数になるまで増加させる．

機器の情報からの世帯属性の推定に向け，これら 3 つの課題を解決する手法を組み合わせた方式を提案する（図 3）．

3. 世帯属性推定の評価

本章では，機器名情報と利用情報の 2 つから，提案手法を用いて世帯属性を推定した際の適合率を求め，世帯属性推定における機器に関する情報の有効性を評価する．

3.1 世帯属性推定の評価内容

3.1.1 世帯属性推定の適合率評価

世帯属性推定として，二値分類と多値分類を行い，それぞれの適合率を評価する．二値分類は，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手法である．二値分類は，世帯に適した広告を提示するサービス等，ある特定の世帯にのみサービスを提供する際に有効な推定である．一方，多値分類は複数の目的変数から，合致する一つの目的変数を推定する手法である．多値分類は，マーケティング等，世帯像を把握する際に有効な推定である．これら 2 つの推定について適合率を評価していく．

3.1.2 説明変数の設計に向けた評価

また，説明変数の種類の多さと粒度が適合率に与える影響に関する知見の獲得を目的に，以下の説明変数を組み合わせた場合の適合率も評価する. これは, 世帯構成以外に，世帯に関する新たな情報の推定を検討する際，種類や粒度を踏まえて，説明変数の設計を行えるようにするためである．

(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
(iii) 機器区分内のメーカごとの台数
(iv) 利用情報

上記(ⅰ)ホームネットワーク内機器の全台数から(ⅱ)機器区分ごとの台数，(ⅲ)機器区分内のメーカごとの台数に従って，情報が細分化されていく．本評価では，(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までのそれぞれ説明変数において各目的変数の適合率を求めていき，説明変数の粒度が適合率に与える影響を検証し，かつ，(ⅳ)利用情報を，(ⅰ)ホームネットワーク内機器の全台数から(ⅲ)機器区分とメーカごとの台数の説明変数に加えて評価することで，種類の多さが適合率に与える影響も検証する．(ⅳ)利用情報は，最大同時利用機器数，時間帯ごとの使用時間合計，オンオフ切り替え回数合計に加工した後の情報である．それぞれの説明変数において各目的変数の推定を実施する．

3.2 説明変数の準備

各世帯のホームネットワークからプロトコル情報を取得する仕組みのない現状において，収集情報を世帯属性と正しく関連付けるため，1,000 世帯に対して web アンケートを実施し，人数と家族構成，所有機器，利用情報の正解データを取得した．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，型番の情報を入力させた．アンケートにてヒアリング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
③TV
④ゲーム機
⑤タブレット
⑥プリンタ
⑦ネットワーク機器（ルータ，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 時間の内 1 回でも使用した場合，使用した旨を申告してもらい，1 時間の内 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 web 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．そのため，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にした．説明変数の一覧は表 1 の通りである．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有無, 機器区分内のメーカごとの台数と有無を変数として用意した．有無情報とは，1 台でも存在していれば数字の 1 を値として持ち，1 台も存在しない場合は数字の 0 を値としてもつ. 例えば, ある単身世帯の男性が, A 社携帯電話 1 台, B 社パソコン 1 台，C 社携帯型ゲーム機 1 台，D 社据え置き型ゲーム機 1 台，E 社ブロードバンドルータが 1 台を所有していた場合の説明変数は表 2 のようになる. 表 2 では, ある世帯が所有している機器のみを行に記載しているが, 実際のテーブルでは全ての機器メーカの行をもっており, 所有していない機器については台数・有無ともに「0」となる．同様に，利用情報においても，ある時間帯に機器を利

表 1　説明変数一覧
Table 1　The list of explanatory variables.

	機器名情報		利用情報
	台数	有無	
全機器	○	×	×
機器区分ごと	○	○	○
機器区分内のメーカごと	○	○	×

○：説明変数として利用　×：説明変数として利用しない

表 2　説明変数の例

Table 2　The examples of explanatory variables (use information is prior to processing).

	機器名情報		利用情報（※）					
			平日（木）			平日（金）		‥
	台数	有無	0:00～1:00	1:00～2:00	‥	0:00～1:00	‥	
全機器	5	-	-	-		-		
①　帯電話	1	1	-	-		-		
A 社携帯電話	1	1	1	0		0		
②パソコン	1	1	-	-		-		
B 社パソコン	1	1	1	1		0		
④ゲーム機	2	1	-	-		-		
C 社ゲーム機	1	1	1	0		0		
D 社ゲーム機	1	1	0	0		0		
⑦ネットワーク機器	1	1	-	-		-		
E 社ネットワーク機器	1	1	1	1		1		

（※）利用情報は加工前の情報

用していた場合,その機器の行における利用情報として「1」とし，利用していなかった場合は「0」となる．

3.3 説明変数に対する提案手法の適用結果

本評価では，木曜日から日曜日までの 4 日間のデータをヒアリングした．2.3.2 節で述べた機器利用情報の丸め処理の結果,(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)時間帯ごとの使用時間合計は，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各々4 つの時間帯について使用時間合計を取得し，説明変数の数は 7 機器区分，8 時間帯分の組み合わせで 56 となった．(3)オンオフ切り替え回数合計は，機器区分ごとに，平日と休日につきオンオフ切り替え回数を取得し,説明変数の数は 7 機器区分の 2 日分の 14 となった．

SMOTE 関数による正解ダミーデータの作成は，正解データの数が少ない，世帯人数が 5 人と 6 人の目的変数においてのみ活用した．

3.4 適合率算出方法

評価では，1,000 世帯分のデータをランダムに 5 グループに分類し，その内 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[8]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検証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 つ選択し，各グループに対して評価用データとし再実験していく．グループ数である，全 5 回の評価結果である適合率の平均を，最終的な適合率として算出する．評価用データを固定せず適合率を求めることで，汎用性の高い評価を行える．

表 3　各目的変数における適合率

Table 3 Precision for each objective variable.

目的変数	二値分類		多値分類		正解データ数
	適合率（%）	期待値（%）	適合率（%）	期待値（%）	
1 人	89.5	50.0	63.6	16.7	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	50.0	64.4	16.7	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．

4. 世帯属性推定の評価結果と考察

本評価では，世帯人数を目的変数として提案手法を適用した．アルゴリズムとして SVM（Support Vector Machine）[9]を用いた世帯属性結果を表 3 に示す．

4.1 世帯属性推定の評価結果

二値分類の適合率を表 3 の左部に，二値分類を拡張した多値分類の適合率を表 3 の右部にそれぞれ示す．二値分類時の期待値は，2 つの値のどちらかを回答するため，期待値は 50.0%となる．一方，多値分類時の期待値は 6 つの値から回答するために，16.7%（=1÷6×100）となる．多値分類は，目的変数から 2 つずつ選択（15 パタン（$=_6C_2$））して二値分類を実施し，ロス関数に基づく復号法[10]を活用して求めた．評価結果は，二値分類の多重平均が 83.7%であり，多値分類の多重平均は 64.4%であった． 二値分類・多値分類ともに，全ての目的変数において推定モデルを構築でき,期待値よりも精度良く推定することができた．本評価により，機器に関する情報から世帯の属性を推測できること検証により確認することができた．

4.2 説明変数の設計に向けた考察

(ⅰ)ホームネットワーク内機器の全台数, (ⅱ)機器区分ご

表 4 各説明変数の組み合わせと適合率（二値分類）の平均
Table 3 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のみ(ⅱ)	62.3
機器区分内のメーカごとの台数のみ(ⅲ)	66.5
利用情報(ⅳ)	56.2
(ⅰ)+(ⅱ)+(ⅲ)+(ⅳ)（4.1 節の結果と同じ）	83.7

との台数，(ⅲ)機器区分内のメーカごとの台数，(ⅳ)利用情報のそれぞれの説明変数で各目的変数に対して推定を実施した際の適合率の平均は表 4 のようになった．所有端末情報において，情報を細分化していくことで，推定の適合率が向上する結果となった．これは，説明変数の情報を細分化していくことで，より目的変数の特徴を表現できるようになったためだと考えられる．また，(ⅰ)ホームネットワーク内機器の全台数から(ⅳ)利用情報まで全てを説明変数に加え，種類を多くした際の推定の適合率は，(ⅲ)機器区分内のメーカごとの台数を説明変数に加えたときよりも適合率が向上している．この結果より，説明変数の粒度を細分化すると適合率が向上し，かつ種類が多くなることによっても適合率が向上することを確認できた．種類を多くすることで推定の適合率が向上するという検証結果も得られたため，現在の所有機器情報と利用情報だけでなく，他の情報も説明変数に追加していくことで推定の適合率がより向上すると考えられる．

説明変数の粒度を細分化することで推定の適合率も向上するという検証結果が得られたため，今後はメーカごとの台数ではなく，シリーズ名のレイヤまで情報を細分化することで推定の適合率が向上すると考えられる．また，利用情報においては，現在時間帯ごとの使用時間合計は，1 日を 6 時間ずつ 4 区分に分類しているが，時間粒度を細分化することでも推定の適合率が向上すると考えられる．一方，粒度を細分化し過ぎると，データ間での類似性が減少し，集合を形成できなくなり，推定モデル構築が不可能になる可能性が発生すると考えられるが，正解データ数を増やすことによって，推定可能な範囲を確保したい．

5. 将来課題

今回の評価では，アンケートにより利用状況を取得したが，実際に機器の利用状況を取得する際は，ICMP echo 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，ICMP echo 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[11]．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．

本評価においては目的変数である世帯属性として，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．例えば，世帯属性情報として，各構成員の年齢，職業がある．構成や各構成員の年齢や表現を表現することにより，さらに世帯像把握の詳細度を高めることが可能になると考えられる．今後，年齢や職業も含めた世帯属性の推定も検証し，有効性を確認していきたい．

また，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．

6. まとめ

本研究では，世帯像の把握に向け，機器名情報と利用情報を学習して世帯属性を推定する手法を提案し，世帯属性の推定に対する機器に関する情報の有効性を評価した．1,000 名のアンケートで機器名情報と利用情報，世帯人数を収集・学習し，推定を実施したところ，二値分類では 83.7%，多値分類で平均 64.4%と，ランダムに選択するよりも精度高く推定でき，世帯属性推定における機器に関する情報の有効性を示すことができた．本提案にて，説明変数の組み合わせを評価する回数を大幅に削減することができ，かつ目的変数となる正解データが少ない場合でも推定モデルを構築することができた．また，本実験において機械学習の説明変数の種類が多くなるほど，粒度が細かくなるほどに適合率が向上することを確認でき，説明変数の設計に有効な知見を獲得することができた．

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) 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.
5) 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.
6) M. W. Browne: "A symptotically distribution-free methods for the analysis of covariance structures," British Journal of Mathematical and Statistical Psychology, vol. 37, Issue 1, pp.62-83, 1984.
7) Nitesh V. Chawla1, Kevin W. Bowyer, Lawrence O. Hall1and W. Philip Kegelmeyer: "SMOTE: Synthetic Minority Over-sampling Technique," Journal of Artificial Intelligence Research 16, pp.321-357, 2002.
8) Geisser, Seymour: "Predictive Inference," CRC Press, 1993. ISBN 0-412-03471-9.
9) Vladimir Vapnik and Corinna Cortes: "Support vector networks," Machine Learning, vol. 20, pp.273-297, 1995.
10) R. W. Hamming; "Error Detecting and Error Correcting Codes," Bell System Technical Journal, Volume.29, Issue 2, pp.147–160, 1950.
11) 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.