奈良女子大学大学院修士論文

食事と健康状態の関連予測のための データマイニングに関する研究

奈良女子大学大学院　人間文化研究科
博士前期課程　情報科学専攻
（学籍番号：I05-010）
李　丹陽

指導教官:城　和貴

平成19年1月

概要

近年，科学進歩に伴い，記憶装置の大容量化が進んでいる．その結果として，蓄積された大量のデータの中から得られる情報は，多種多様かつ複雑である．そのため，従来の統計解析手法では扱うことが難しいデータや，様々な形式のデータベースから，有用な情報を取り出す必要がある．このための技術として，データマイニングが注目されている．

生活習慣病の予防，エネルギー・栄養素欠乏の予防，過剰摂取による健康障害の予防，健康維持・増進などを図るための適切な栄養摂取量，望ましい食生活のあり方を追求するため，健康的な食事摂取プランの開発に関する研究がある．また，食物摂取頻度調査による，食品や栄養の摂取量から日常の食事の内容を評価する食物摂取頻度調査を把握する研究が盛んである．しかし，これらは食品と健康状態の直接的な関連に関する研究ではない．食事による摂取エネルギー，運動によるエネルギー，睡眠時間，飲酒量，喫煙量など生活習慣データと，血圧，体重，体脂肪率など健康状態データに関する相関ルール解析を行う健康データマイニングシステムの開発研究もある．しかし，食品の摂取量，睡眠時間，飲酒量，喫煙量など生活習慣データが必要なため，被験者の負担が大きい．この負担を減らすためには，大まかな食品の摂取と健康状態のみのデータから，特徴を発見し，健康状態の把握や管理をすることが考えられる．このことにより，より日常的に簡単に食事と健康の関係を知ることができると考えられる．そこで，摂取した食品と健康状態のデータに対してデータマイニングを適用することによって，摂取した食品と健康状態の関連について人間の先入観を介入させず発見するためのシステムを構築すべきである．健康状態を知るためのパラメータとして，同研究室では，小松原が排泄物の形状から健康状態を推論するための一手法に関する研究を行っている．一方，本研究では，摂取した食品と健康状態に対するデータマイニングの開発を行う．これにより，健康状態の管理に役立つ指標を作る．このデータマイニングの性能を調べるために，実験を行う．なお，小松原が開発中の部分は，まだ未完成であるため，本論文の実験では，被験者が直接健康状態を入力したデータを利用することとする．

そこで，本論文では，摂取した食品と健康状態のデータに対してデータマイニングを適用することで，摂取した食品と健康状態の関連について人間の先入観を介入させず発見し，健康状態の管理に役立つ指標を作るために，食事と健康状態の関連を調べる手順を提案する．また，データマイニングを用いて食事と健康状態の相関ルールを発見するための実験を行う．

データマイニングとは，統計学，パターン認識，人工知能などのデータ解析の技法を用い，大量のデータを分析し，隠れた関係性や意味を見つけ出す技術である．データベースに蓄積された大量のデータから相関ルールを抽出する技術を相関ルール抽出，あるいは相関ルール

分析という．自動的にデータベースから価値のある相関ルールを効率的にかつ漏れなく発見する方法として，アプリオリアルゴリズムがある．アプリオリアルゴリズムは,「長さ k の頻出でないパターンを含む長さ $k+1$ のパターンは頻出でない」という理論の元で，頻出パターンを抽出するアルゴリズムである．そして，抽出された相関ルールを評価するため，カイ 2 乗検定を行うことによって，明らかに価値のない相関ルールをとり除くことができる．データマイニングで良く用いられる代表的な手法として，決定木がある．決定木とは，データベースに蓄積された複雑な事象を相関ルールを用いて表現し，根または分岐ノードが属性テスト，枝が分割テストの結果，葉ノードがクラスラベルあるいはクラス分布を表すような木構造である．本論文では，アプリオリアルゴリズムを用い，提案した食事と健康状態の関連に関する調べる方法を，データについて適用する実験を行う．

今回の実験では，共同研究の相手の都合により，任天堂 DS のソフトに含まれる 198 品目のレシピを分析対象に選ぶ．実験データとして，20 代の女性 32 名をデータの対象者とし，1ヶ月に食べたレシピ履歴データスクリプトによって生成する．データの中から，日付，食べたレシピ，便通状態の 3 つのデータを切り取って使用した．便通状態は便秘の場合のみ実験を行う．蓄積されたデータセットに対して，アプリオリアルゴリズムを用いてデータマイニングを行う．今回，便秘の前日の 1 日分のレシピのみからなるデータ集合に着目して単体および組み合わせでデータマイニングを行う．また，便秘の前日だけのデータが必ずしも便秘に影響しているとは限らないと考え，便秘の前の複数日間分のレシピからなるデータ集合に着目して組み合わせてデータマイニングを行う．

本論文では，データマイニングを用いて効果的に健康に良い食べ物，健康によくない食べ物の特徴を発見することで，食べ物の選択および健康状態の管理に役立つ指標を見出すことを目的とし，データマイニングを用いて食事と健康状態の関連の調べに関する方法を提案し，実験を行った．実験の結果から，前日に食べると便秘になる可能性が高い食事の組み合わせを見出す．また，便秘の前日だけのデータが必ずしも便秘に影響しているとは限らないことも示す．

キーワード：データマイニング，相関関係

目 次

図 目 次

表 目 次

第1章　はじめに

近年，科学進歩に伴い，記憶装置の大容量化が進んでいる．その結果として，蓄積された大量のデータの中から得られる情報は，多種多様かつ複雑である．そのため，従来の統計解析手法では扱うことが難しいデータや，様々な形式のデータベースから，有用な情報を取り出す必要がある．このための技術として，データマイニングが注目されている．データマイニングを用いることにより，集めたデータからなんらかの知見を発見することが期待される．人間は，良好な健康状態を得るために，摂取する食べ物に関する情報を必要とする．食物と健康の関連を知るために，被験者のアンケート結果をもとに，データマイニングを行う．

本論文では，食事と健康状態の関連を調べるための手順を提案する．また，データマイニングを用いて食事と健康状態の相関ルールを発見する実験を行う．

本論文では以下の形で構成される．第2章で関連研究について述べる．第3章でまず，データマイング，相関ルールおよびアプリオリアルゴリズムについて述べる．そして，決定木およびクラスタリングについて説明する．第4章で実験環境について述べる．第5章で提案方法を実装し，評価手法と結果について述べる．第6章で実験結果を考察し，最後に第7章でまとめ．

第2章　関連研究

生活習慣病の予防，エネルギー・栄養素欠乏の予防，過剰摂取による健康障害の予防，健康の維持・増進などを図るための適切な栄養摂取量，望ましい食生活のあり方を追及するため，健康的な食事摂取プランの開発 [1]，食事療法についての提案する研究 [2]，健康的な食生活習慣形成を目指した食事摂取基準に関する研究 [3]，食事調査法の開発 [4]，食事摂取基準の活用に関する実践的研究 [5]，食物摂取頻度調査による，食品や栄養の摂取量から日常の食事の内容を評価する食物摂取頻度調査を把握する研究 [6] が盛んである．しかし，これらは食品と健康状態の直接的な関連に関する研究ではない．食事による摂取エネルギー，運動によるエネルギー，睡眠時間，飲酒量，喫煙量など生活習慣データと，血圧，体重，体脂肪率など健康状態データに関する相関ルール解析を行う健康データマイニングシステムの開発研究 [7] もある．しかし，食品の摂取量，睡眠時間，飲酒量，喫煙量など生活習慣データが必要なため，被験者の負担が大きい．この負担を減らすためには，大まかな食品の摂取と健康状態のみのデータから，特徴を発見し，健康状態の把握や管理をすることが考えられる．このことにより，より日常的に簡単に食事と健康の関係を知ることができると考えられる．そこで，摂取した食品と健康状態のデータに対してデータマイニングを適用することによって，摂取した食品と健康状態の関連について人間の先入観を介入させず発見するためのシステムを構築すべきである．健康状態を知るためのパラメータとして，同研究室では，小松原が排泄物の形状から健康状態を推論するための一手法に関する研究を行っている．一方，本研究では，摂取した食品と健康状態に対するデータマイニングの開発を行う．これにより，健康状態の管理に役立つ指標を作る．このデータマイニングの性能を調べるために，実験を行う．なお，小松原が開発中の部分は，まだ未完成であるため，本論文の実験では，被験者が直接健康状態を入力したデータを利用することとする．

第3章　データマイニング

本章では，データマイニングについて説明する．データマイニングとは，統計学，パターン認識，人工知能などのデータ解析の技法を用い，大量のデータ分析し，隠れた関係性や意味を見つけ出す技術である [8]．データマイニングの定義としては，明示されておらず今まで知られていながったが，役立つ可能性があり，かつ自明でない情報をデータから抽出することである [9]．データベースに蓄積された大量のデータから相関ルールを抽出する技術を相関ルール抽出，あるいは相関ルール分析という．

3.1 節では，相関ルールについて述べ，サポートおよび確信度について述べる． 3.2 節では，アプリオリアルゴリズムの生成例について説明する．さらに，アプリオリアルゴリズムの利点および欠点にいて記述する．3.3 節では，相関ルールの評価基準としてカイ 2 乗検定について述べる．3.4 節では，データマイニングで良く用いられる代表的な手法決定木の分割テストおよび決定木の構築アルゴリズムについて説明する． 3.5 節では，クラスタリングについて述べる．

3.1　相関ルール

この節では，まず相関ルールについて述べる．次に，サポートおよび確信度について述べる．

相関ルールとは，ある事象が発生すると別の事象が発生するといったような，同時性や関係性が強い事象の組み合わせ，あるいはそうした強い事象間の関係のことである．スーパーマーケットで売られている商品をアイテムと呼び，顧客が購入したアイテムリストをトランザクションと呼ぶ．例えば,「パンを購入した顧客のうち，85%が牛乳も購入しており，この 2 種の商品すべてを購入した顧客は全顧客の 6%である.」というようなことが得た場合，

$\{$パン$\} \Rightarrow \{$牛乳$\}$: $sup = 6\%, conf = 85\%$

という式で表現できる．一般に X，Y を商品の集合として（この例では $X =$ $\{$パン$\}$，$Y =$ $\{$牛乳$\}$），

$X \Rightarrow Y$

と表現されることを相関ルールと呼ぶ [10]．ここで，X を前提部と呼び，Y を結論部と呼ぶ．データベース D 中から全アイテムの集合を I とし，その部分集合をアイテムセットと呼ぶ．また，与えた最小サポート以上のサポートをもつアイテム集合を頻出アイテム集合と呼ぶ．ここで各トランザクション T は I の部分集合である．D 中の全トランザクションの

うち，アイテムセット X を含むトランザクションの割合を X のサポートといい，$sup(X)$ と表記する．相関ルール $(X \Rightarrow Y)$ のサポートは $sup(X \cup Y)$ で，確信度 $conf(X \Rightarrow Y)$ は $sup(X \cup Y)/sup(X)$ で定義される．最小サポート以上のサポートをもつ．

どのアイテムを組み合わせれば価値のある相関ルールができるかを調べる必要がある．そこで，自動的にデータベースから価値のある相関ルールを漏れなく効率的に発見すべきである．そこで，IBM アルマデン研究所の R.Agrawal によって提案されたアプリオリアルゴリズムの手法は，世界初の本格的なデータマイニングシステムである [11].

3.2 アプリオリアルゴリズム

この節では，アプリオリアルゴリズム [10] の生成例について説明する．さらに，アプリオリアルゴリズムの利点および欠点について述べる．

アプリオリアルゴリズムは,「長さ k の頻出でないパターンを含む長さ $k+1$ のパターンは頻出でない」という理論の元で，頻出パターンを抽出するアルゴリズムである [12]．相関ルールの生成に必要なデータ構造を主記憶内につくることによって，効率的にすべての相関ルールを発見することができる．

アイテム集合のサポートを計算するために，データベースをスキャンし，アイテム集合のトランザクションの数を数えなければならない．一回のスキャンでいくつかのアイテム集合のサポートをまとめて計算する．また，要素の少ないアイテム集合からそれぞれのサポートを調べ，あるアイテム集合のサポートが最小サポートより小さいと分かったら，それを含むようなアイテム集合も決して頻出集合ではないので候補のアイテム集合の生成をしないようにする．アプリオリアルゴリズムでは，k 回目のスキャンで要素数 k のアイテム集合のサポートを求める．

ここで，C_k は要素数 k のアイテム集合の候補のアイテム集合とし，L_k は要素数 k の頻出アイテム集合の集合とする．アプリオリアルゴリズムは以下のようになる．

1. 要素数のアイテム集合の候補アイテム集合を C_1（相関ルールとして抽出される候補）とする．全データベースを検索して各候補アイテム集合 C_1 の出現回数をカウントし，サポートを計算する．

2. 各候補アイテム集合 C_1 について，定めた基準である最小サポートを満たすサポートを持つ候補アイテム集合を頻出アイテム集合 L_1 とする．

3. 頻出アイテム集合 L_1 同士の組み合わせを新しい候補アイテム集合 C_2 として出現回数をカウントし，サポートを計算する．

4. ステープ 2 と 3 の処理を k 回繰り返し，候補アイテム集合 C_k が空になるまで続ける．

図 3.1 に示すデータベースの例を使って説明する．図 3.1 にある最初の表のデータベースにおいて，各行がトランザクションを表す．この例では，最小サポート 50%とする．この

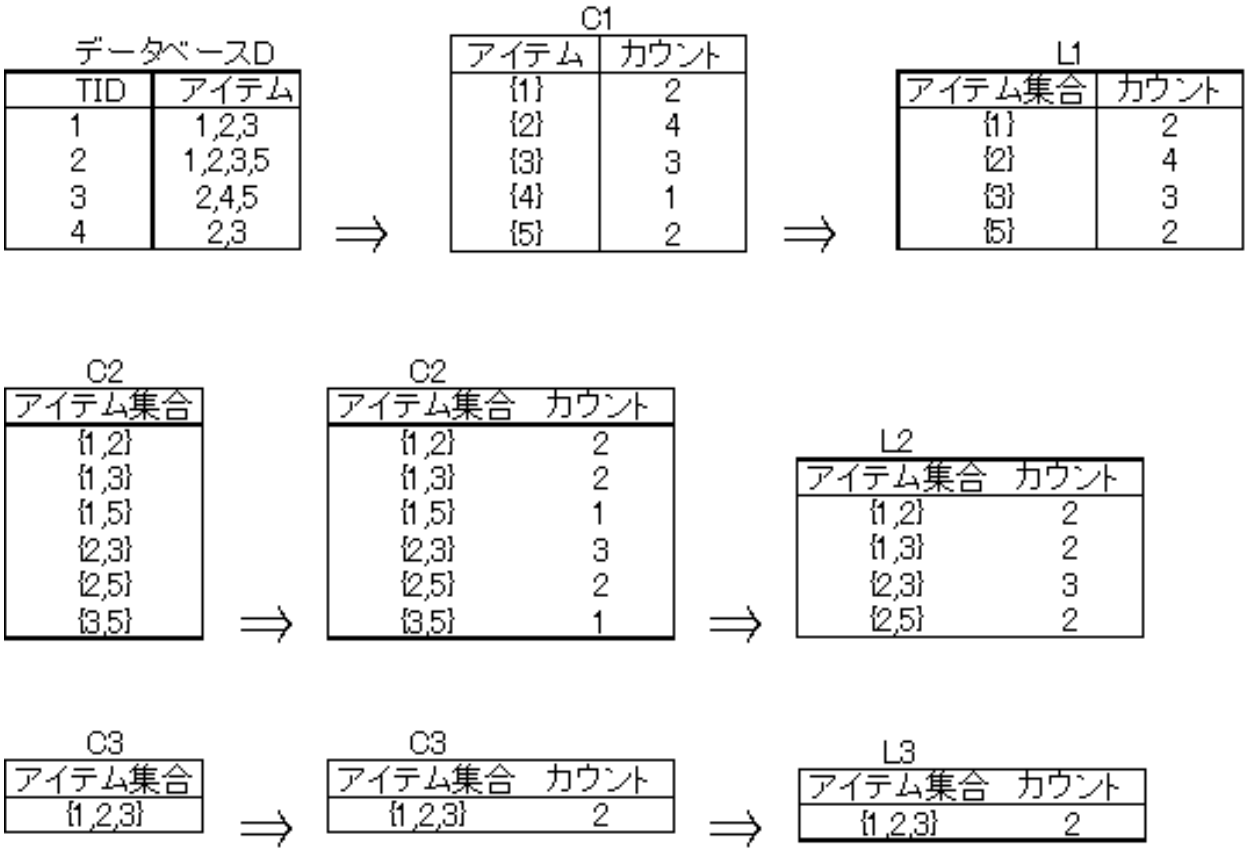

図 3.1: 候補アイテム集合と頻出集合の生成例

データベースには 4 つのトランザクションがあるから，2 つ以上のトランザクションに含まれるアイテム集合が頻出集合となる．まず，すべての要素数 1 のアイテム集合を C_1 とし，データベースをスキャンし，それぞれのサポート数をカウントする．{4} は一つのトランザクションにしか出現しないため除かれ，残りのアイテム集合を頻出アイテム集合 L_1 の要素数とする．次に，L_1 から要素数 2 の候補アイテム集合 C_2 をつくり，データベースをスキャンしてそれぞれのサポート数をカウントする．{1,5}, {3,5} が 2 つ共に最小サポートを満たさないため除かれ，それ以外のアイテム集合を L_2 の要素数とする．頻出アイテム集合 L_2 から生成される候補アイテム集合は {1,2,3} のみであり，その出現回数を数え，L_3 をつくる．そして，L_3 からは要素数 4 の候補アイテム集合をつくれないため，このアルゴリズムが終了する．

アプリオリアルゴリズムの利点として，次のことが考えられる．

アプリオリアルゴリズムでは，候補アイテム集合をつくる時に，その部分集合のすべてが 1 つ前の頻出アイテム集合に出現するもののみを抽出する．これは部分集合の 1 つでも最小サポートを満たさないものあるアイテム集合は，当然最小サポートも満たさないという考え

	t/yes	t/no	合計
m/yes	50	15	75
m/no	30	5	35
合計	80	20	100

表 3.1: 牛乳と紅茶を飲むデータ

に基づく．このため，生成必要がない候補アイテム集合が大幅に削減することが出来る．これは今回の実験がアプリオリアルゴリズムの手法を用いる理由である．

アプリオリアルゴリズムの欠点として，以下のことが考えられる．

アプリオリアルゴリズムにおける候補アイテム集合の中で起こりうるすべてのアイテムの組み合わせを含んでいる．そのため，候補アイテム集合が莫大な数になってしまう可能性がある．そして，抽出するアイテム集合の長さが長くなればなるほど，必要となる候補アイテムが指数的に増大するため，記憶容量が多く必要となってしまう．また，データベースのスキャン回数が多くなってしまうため，記憶容量が大きく必要になってしまう．

3.3 相関ルールの評価基準

この節では，相関ルールの評価基準としてカイ 2 乗検定 [11] について述べる．

相関ルールの価値の評価基準として，これまで確信度とサポートを用いると述べた．しかし，高い確信度を持ちながらも強い相関をもたない場合がある．

例えば，表 3.1 のようなデータがある．紅茶を飲む人 (t/yes) と飲まない人 (t/no) の数を表の列に表し，牛乳を飲む人 (m/yes) と飲まない人 (m/no) の数を表の行に表している．相関ルールを $\{m/yes\} \Rightarrow \{t/yes\}$ とする．相関ルールの確信度は $50/75 = 67\%$である．しかし，紅茶を飲む人の全体の割合 ({t/yes} のサポート) は $80/100 = 80\%$であり，この相関ルールの確信度より高い．つまり，確信度が高いにもかかわらず，紅茶を飲む人はむしろ牛乳をあまり飲まないということとなる．

上の例から相関ルールの価値の評価基準としてサポートと確信度による評価のみでは必ずしも適切とはいえない．そこで抽出された相関ルールの価値を適切に評価するため，確信度とサポート値を評価基準として用いるだけでなく，それに加えて更にカイ 2 乗検定も行う．

カイ 2 乗検定とは，ある仮説のもと二つの事象を調査し，統計的な有意性があるかどうかを判定することである．例えば，顧客が選ぶ商品の週ごとの変化が，意味のあるレベルの変動しているかを判定する時に利用できる．

統計学では，表 3.1 のような表を分割表と呼ぶ．分割表がランダムなサンプルから得られると仮定できる時，分割表から とよばれる独立性を使った検定を応用して，価値のない相関ルールを取り除く方法がカイ 2 乗検定である．

相関ルール $X \Rightarrow Y$ とする．X，Y，$X \cup Y$ のサポートをそれぞれ S_X，S_Y，S_{XY} とし，

トランザクションの総数を N とする．ここで X と Y が独立し，同じトランザクション内に含まれるのが単なる偶然であると仮定する．カイ 2 乗の検定量 T_{dep} は次のようになる．

$$T_{dep} = N \frac{(S_{XY} - S_X)^2}{S_X S_Y (1 - S_Y)(1 - S_X)} \tag{3.1}$$

検定量は自由度 1 のカイ 2 乗分布に従うことが知られている．カイ 2 乗の検定量 T_{dep} の値が 0 に近ければ X と Y はお互いに独立であり，大きければ相関が強いといえる．そこで，ある有意水準 α を定め，$T_{dep} < {x_1}^2(\alpha)$ であれば X と Y が独立であると見なし，相関ルール $X \Rightarrow Y$ が発見されたのは単なる偶然であるから，価値がないとして捨てる．

3.4 決定木

この節では，データマイニングで良く用いられる代表的な手法決定木について述べる．まず，決定木の分割テストについて説明する．次に，決定木の構築アルゴリズムについて述べる．

データベースから抽出された相関ルールは，様々なデータ分析に応用することができる．各相関ルールをデータ分類，値の予測などに応用する代表的な手法として決定木がある．決定木は木構造の特別な形である．決定木とは，データベースに蓄積された複雑な事象を相関ルールを用いて表現し [13]，根または分岐ノードが属性テスト，枝が分割テストの結果，葉ノードがクラスラベルあるいはクラス分布を表すような木構造である．クラスとは，ある事例がどういう集合に属するかを表す [14]．頂点ノードの分割テストであるかどうかで，下位ノードに分類される．こうした分類を繰り返すことによって，最終的にいずれかの終端ノードに分類される．決定木は，知識・法則を頂点ノードから終端に至るまでの，分割テストの IF=THEN ルールとして簡単に表現することができる．終端ノードのラベルは，IF-THEN ルールの結論部となる．

3.4.1 決定木の分割テスト

表 3.2 の例を使って説明する．このデータベースのように決定木構築に利用されるデータを学習データと呼ぶ．データベースの中の属性のうち,「商品 A」のように決定木を構築する時に，IF-THEN ルールの結論部に現れる属性を目的属性と呼び，分割テストで利用される「年齢」,「性別」と「高級商品 1 所有」の属性を条件属性と呼ぶ．

この学習データから抽出される相関ルールとして以下のものとなる（表 3.3 と 3.4 に参照）．

- ルール Y1
  高級商品 1(Y1) を持っている人は商品 A(X) をよく購入する

年齢	性別	高級商品 1 所有	商品 A
女	20 代	Y	0
女	10 代	Y	1
男	30 代	N	1
男	20 代	N	1
女	10 代	Y	0
男	30 代	N	1
女	20 代	N	0
男	20 代	Y	1

表 3.2: 学習データの例

	高級商品 1 所有	高級商品 1 なし	合計
商品 A ○	2	3	5
商品 A ×	2	1	3
合計	4	4	8

表 3.3: ルール Y1

- ルール Y2

  男性 (Y2) は商品 A(X) をよく購入する

ここで,「商品 A」の値を予測することについて考える．例えば，高級商品 1 を持っている女性の場合に対して「商品 A」の値を予測すると，ルール Y1 で判断するならば「商品 A」を購入する，ルール Y2 ならば「商品 A」を購入しないと予測することができる．

このように，相関ルール同士を比較する指標としては扱いにくいため，単一の評価値をもつ別の評価関数を利用することが多い．よく利用される評価関数としては，相互情報量がある．

あるデータ集合の事象 X に関するあいまいさは以下の式で定義されるエントロピー関数で測ることができる．

$$H(S) = H(X) = -\sum_{i=1}^{k} p_i \log_k p_i$$

ここで，p_i は X の k 個ある事象 $a_i (1 \leq i \leq k)$ の起こる確率とする．

ルール Y1 と「商品 A」との関連は，表 3.3 のような結果が得られるとする．エントロピー関数値は

$$H(X) = -5/8 \log 5/8 - 3/8 \log 3/8 = 0.95$$

である．ルール Y1 を満たす場合のエントロピーは

	男	女	合計
商品 A ○	4	1	5
商品 A ×	0	3	3
合計	4	4	8

表 3.4: ルール Y2

$$H(X|Y1 = yes) = -2/4 \log 2/4 - 2/4 \log 2/4 = 1$$

である．同様に，ルール Y1 を満たさない場合

$$H(X|Y1 = no) = -3/4 \log 3/4 - 1/4 \log 1/4 = 0.8113$$

となる．平均エントロピーの関数値は

$$H(X|Y1) = 4/8 \times H(X|Y1 = yes) + 4/8 \times H(X|Y1 = no) = 0.9057$$

までに減少する．このエントロピー関数値の減少量は

$$H(X) - H(X|Y1) = 0.95 - 0.9057 = 0.0443$$

となる．これはルール Y1 の「商品 A」に関する相互情報量である．

ルール Y2 と「商品 A」との関連として，表 3.4 のような結果が得られる．エントロピー関数値は

$$H(X) = -5/8 \log 5/8 - 3/8 \log 3/8 = 0.95$$

である．ルール Y2 を満たす場合のエントロピーは

$$H(X|Y2 = yes) = -4/4 \log 4/4 - 0/4 \log 0/4 = 0$$

である．同様に，ルール Y2 を満たさない場合

$$H(X|Y2 = no) = -1/4 \log 1/4 - 3/4 \log 3/4 = 0.8113$$

となる．平均エントロピーの関数値は

$$H(X|Y2) = 4/8 \times H(X|Y2 = yes) + 4/8 \times H(X|Y2 = no) = 0.4057$$

までに減少する．このエントロピー関数値の減少量は

$$H(X) - H(X|Y2) = 0.95 - 0.4057 = 0.5443$$

となる．これはルール Y2 の「商品 A」に関する相互情報量である．ルール Y2 の方が相互情報量が大きい．このようにして求めた相互情報量は,「商品 A」に関するルール同士を比較するのに利用され，相互情報量の大きいルールで予測するほうがよい．この場合では，ルール Y2 つまり男性は商品 A をよく購入するということで予測する．

3.4.2 決定木の構築アルゴリズム

決定木を作成する時，どの条件属性に対し，どのような分割テストをどのような順番で適用するかによって，構築される木の大きさが決まる．一般的には，分割テストが少ないほうが望ましいとされている．しかし，木の高さ（根から葉へのパス長をすべての葉について合計したもの）を最小の木を構築するのは NP 困難であることが分かっている [15]．解決方法として，再帰的に相互情報量などに基づいたバックトラックを行わない最適分割テストにより分割していく貪欲アルゴリズム (greedy algorithm) がある [16][17].

基本的な決定木構築のアルゴリズムは以下のようになる．

メインルーチン (main)

1. データベース中の全学習データ D を読み出す
2. SPLIT(D)

サブルーチン (SPLIT(データ集合 D))

1. IF（D が分割終了条件を満たす）THEN 終了
2. 各カテゴリ型属性に対し最適な分割テストを探す
3. 各数値属性に対し最適な分割テストを探す
4. (2),(3) で見つかったすべての最適分割テストのうち，最も目的関数値のよいテストでデータ集合 D を D_1 と D_2 に分割する
5. SPLIT(D_1)

6. SPLIT(D_2)

このアルゴリズムの終了条件は以下のようにとなる

- データ集合 D の目的属性値がすべて同値か，一つの目的属性値の存在比率が十分大きい
- 条件属性上で定義可能な分割ルールでは，データ集合 D をこれ以上分割できない
- データ数 $|D|$ が全学習データ数に対して十分小さい

3.4.3 決定木の調整

決定木を目的属性の値を予測するツールとして利用する時，木の精度が高くなるように木の大きさを調整する必要がある．クロスバリデーション (cross validation) は決定木の予測精度を測る代表的な手法の一つである．クロスバリデーションは，目的属性値の分かるデータを 2 分割し，一つを決定木を構築するための学習データとして，もう一方を構築された決定木の精度を測る検証データとして，互いに構築と検証を行うことである．

クロスバリデーションの検証として，N フォールドクロスバリデーションがある．以下の手順で行われる．

1. 目的属性値の分かるレコードを，ほぼ大きさの等しい N 個の部分データ集合にランダムに分割する
2. N 個の部分データ集合のうち，$N-1$ 個を選んで併合したものを学習データとし，それを用いて決定木を構築する
3. 残りの 1 個を検証データとして，(2) で作成した木の精度を求める．
4. (2),(3) を各部分データ集合が 1 回検証データになるように計 N 回行い，精度の平均を求める

決定木の精度：

$$\text{エラー率} = \frac{\text{予測を誤ったデータ数}}{\text{検証データ数}}$$

生成された決定木は，分岐が多くなりすぎることがある．ノイズデータを含むデータや，例外的な値や誤りに対しても適合しているかもしれないため，結果として予測精度が悪くなってしまう．このように，分析に用いる学習データの例外的な値や誤りに対して適合しすぎた状態を過学習 (overfitting) と呼ぶ．この過学習を避けて正確な予測モデルを構築するために，枝刈りを行う．木の構築を途中で，過学習であるかを判断し，過学習ならば，そこでデータの分割を終了することを事前枝刈りと呼ぶ．また，木を構築し，その木の過学習である部分木を後で取り除くことを事後枝刈りと呼ぶ．

事前枝刈りは，各ノードの最適分割テストによるデータ分割前後における，カイ 2 乗検定，相互情報量などの情報から，そのノードのデータ分割が精度を事前に予測して，データ分割を続けるか終了するかを判断する．

事後枝刈りは，決定木を過学習状態になるまで十分大きく構築し，その状態から過学習となる部分のノードを削除する．一般的には，決定木ではノードの目的属性値がすべて等しくなるか，条件属性値では不可分な状態になるまでの木をまず構築する．事前枝刈りに比べて，コストが余計に必要となるが，事後枝刈りによる決定木のほうが，予測精度が高くなることが多く，現在はこの手法による決定木の調整を行うシステムが多い．

決定木の利点は，ルールを容易に自然言語や SQL に翻訳可能である [18]．そして，データの異常値や分布の歪みに対して頑健である．また，入力変数が欠損していても学習可能であることが挙げられる．

決定木の欠点は，入力変数に連続値が多い問題では性能が落ちることである．また，時系列データを扱う場合はデータの整備が大変であることが挙げられる．

3.5 クラスタリング

前節で述べた決定木は，データを目的属性ごとに，分類するという種分け作業である．しかし，特に目的属性を指定しない場合でもデータを種分けしたいという要求は頻繁に生じる．このような種分け作業を一般にクラスタリングと呼ぶ．決定木分析は，与えられたクラス情報に合うように分類基準を作る．すなわち，教師付き学習であるのに対し，クラスタリングは，分類すべきクラス（目的属性）自身が分からない，すなわち，教師なし学習と呼ばれる．分割後の部分集合をクラスタと呼ぶ．クラスタリング手法は大きく分けて二つある．最短距離法などの階層的手法と，k-means などの分割最適化手法である．

まず，階層的手法の凝集型について述べる．この手法は，N 個の対象からなるデータが与えられたとき，1 個の対象だけを含む N 個のクラスタがある初期状態を作る．この状態から始めて，対象 x_1 と x_2 の間の距離 $D(x_1, x_2)$ からクラスタ間の距離 $D(C_1, C_2)$ を計算し，最も距離の近い二つのクラスタを逐次的に併合する．すべての対象が一つのクラスタに併合されるまで繰り返すことによって，階層構造ができる．クラスタ C_1 と C_2 の距離関数 $D(C_1, C_2)$ の違いにより以下のような手法がある．

最短距離法

$$D(C_1, C_2) = \min_{x_1 \in C_1, x_2 \in C_2} D(x_1, x_2)$$

最長距離法

$$D(C_1, C_2) = \max_{x_1 \in C_1, x_2 \in C_2} D(x_1, x_2)$$

群平均法

$$D(C_1, C_2) = \frac{1}{n_1 n_2} \sum_{x_1 \in C_1} \sum_{x_2 \in C_2} D(x_1, x_2)$$

ウォード法

$$D(C_1, C_2) = E(C_1 \cup C_2) - E(C_1) - E(C_2)$$

ただし $E(C_i) = \sum_{x \in C_i} (D(x, c_i))^2$

ウォード法は，各対象から，その対象を含むクラスタのセントロイドまでの距離の 2 乗の総和を最小化する [19].

次に，分割クラスタリングについて述べる．分割最適化手法は，非階層的手法である．代表的な k-means（k-平均法）は，セントロイド c_i（クラスの重心点）をクラスタの代表点とし，

$$\sum_{i=1}^{k} \sum_{x \in C_i} (D(x, C_i))^2$$

の評価関数を最小化するように k 個のクラスを分割する．

最初に適用するクラスタリング手法は一般に以下のようになる．まず，対象が属性ベクトルで与えられる場合，計算量が k-means 法は $O(NK)$ に対し，階層的手法は $O(N^2)$ なので，k-means 法を用いるほうがよい．対象間の距離だけが与えられる場合は，群平均法を適用する [20].

第4章　実験環境

本章では，実験の目的，実験で用いたデータおよび実験を行う前提条件について述べる．

4.1　目的

まず，健康によい食べ物，健康によくない食べ物の特徴を発見する題材として，共同研究の相手の都合により，任天堂 DS のソフトに含まれるレシピを分析対象に選ぶ．人間は，自分の健康状態を良くするため，どの食べ物を摂取すれば良いかを知りたい．食べ物と健康の関連に基づき，健康に良い食べ物および健康に良くない食べ物の特徴を発見できるならば，選択することも簡単になる．

4.2　実験用のデータ

今回，20 代の女性 32 名をデータの対象者とする．形式は図 4.1 に参照する．32 名の人それぞれに id を付け，食べた日の日付を入力し，対応するところにチェック入れる．何も食べていないなら，何も食べていないところのみチェックを入れる．

任天堂 DS のソフトに含まれるレシピは，全部で 198 品目である．これらは，肉類，魚介，野菜，豆腐，ご飯，めん，汁物，その他とおやつの 9 種類に分けられ，それぞれレシピ番号 1 から 198 までを付けられる（表 4.1 から表 4.9 に参照）．

肉類		
梅しょうゆのさっぱり豚肉	ハンバーグ	鶏ささ身のにんにくじょうゆ焼き
牛タンの七味焼き	棒棒鶏 (バンバンジー)	鶏肉とカシューナッツのいためもの
牛肉ときのこのすだちじょうゆ和え	ビーフシチュー	鶏のから揚げ
牛肉と野菜のみそいため	ひき肉とニラのカレーいため	鶏のチーズかつ
牛肉とレタスのオイスターソースいため	ひき肉と春雨の中華風煮込み	鶏の照り焼き
牛肉のアスパラチーズ巻き	ひとくちステーキ	鶏レバーのしょうが煮
牛肉のしぐれ煮	豚しゃぶのさんしょうダレかけ	とんかつ
牛肉のしょうが風味いため	豚肉のキムチいため	焼きギョーザ
牛肉のにんにくバター焼き	豚肉のしょうが焼き	野菜たっぱり焼肉
水ギョーザ	豚肉の薬味おろし煮	ゆで豚の中華風サラダ
すき焼き	豚の角煮	冷静牛しゃぶ
酢豚	豚ヒレ肉のひとくちステーキ	レバニラいため
チンジャオロース	蒸し鶏の辛味ソース	ローストチキン

表 4.1: データセットレシピ (肉類)(レシピ番号 1-39)


id

日付 年 月 日

健康状態

普通 良い 悪い

便がでなかった 「かたい」便がでた 「やわらかい」便がでた 「水状の」便がでた

何も食べていない

食べたメニューの個数を選択してください

肉

梅しょうゆの さっぱり豚肉	牛タンの七味焼き	牛肉ときのこの すだちじょうゆ和え
牛肉と野菜のみそいため	牛肉とレタスの オイスターソースいため	牛肉の アスパラチーズ巻き
牛肉のしぐれ煮	牛肉の しょうが風味いため	牛肉の にんにくバター焼き
水ギョウザ	すき焼き	酢豚
チンジャオロース	鶏ささ身の にんにくじょうゆ焼き	鶏肉とカシューナッツの いためもの
鶏のから揚げ	鶏のチーズカツ	鶏の照り焼き
鶏レバーのしょうが煮	とんかつ	ハンバーグ
棒棒鶏(バンバンジー)	ビーフシチュー	ひき肉とニラの カレーいため
ひき肉と春雨の 中華風煮込み	ひとくちステーキ	豚しゃぶの さんしょうダレかけ
豚肉のキムチいため	豚肉のしょうが焼き	豚肉の薬味おろし煮
豚の角煮	豚ヒレ肉の ひとくちステーキ	蒸し鶏の辛味ソース
焼きギョーザ	野菜たっぷり焼き肉	ゆで豚の中華風サラダ
冷製牛しゃぶ	レバニラいため	ローストチキン

魚介

あさりの酒蒸し	あじの塩焼き	あじの和風ハンバーグ
いかとえのきの しそいため	いかともやしの からし酢じょうゆ	いかのトマト煮込み
いかの薬豆和え	いわしのしょうが煮	うざく
えびと春雨の タイ風サラダ	えびフライ	かきと卵の ピリ辛いため
かきのみそ焼き	かきフライ	かつおのつくり みそダレ和え
かれいの煮つけ	さけの和風ムニエル	さばのおろし煮
さばの竜田揚げ	さがのみそ煮	さわらの辛味焼き
さんまの塩焼き	塩ざけと野菜のレモン風味いため	スモークサーモンと玉ねぎのポン酢しょうゆ
たいの 和風ごまダレサラダ	たこときゅうりの酢の物	たことしめじの 辛味しょうゆかけ
たこと春菊のキムチ和え	たことトマトの地中海風サラダ	たこのしょうがじょうゆ焼き
たこのにんにく酢みそ和え	ツナとポテトのからしマヨネーズ	はまちのそぎ造りかわりダレ
ぶり大根	ぶりの照り焼き	ほたて貝の香りバター焼き
ほたて貝柱のしょうゆ和え	ほたてと大根のからしマヨネーズ和え	まぐろのたたきとトマト モッツァレラのサラダ

野菜

オクラと鶏ささみの 梅がつお和え	ガーリックポテト	かぶの 鶏そぼろあんかけ
かぼちゃの ガーリック風味焼き	かぼちゃの 煮物	きゅうりの とうがらし漬け
きゅうりもみ	きんぴらごぼう	ゴーヤチャンプルー
ごぼうと鶏ささ身の サラダ	小松菜と油揚げの 煮びたし	里いもとたこの煮物
さやいんげんのごま和え	ししとうがらしと ちりめんじゃこのいり煮	新じゃがの煮物
即席ピクルス	大根サラダ じゃこドレッシング	筑前煮
中華風即席漬け	なす・トマトと ベーコンのチーズ焼き	なすとトマトの 冷製サラダ
なすとひき肉の みそいため	なすの揚げ煮	なめことえのきの おろし和え
肉じゃが	白菜の刻み漬け	八宝菜
ピーマンの焼きびたし	フライパンで作る じゃがいものグラタン風	ブロッコリーとほたての チーズ焼き
ベジタブルシチュー	回鍋肉片（ホイコーロー）	ほうれん草とえびの 鶏いため
ほうれん草とかきの グラタン	ほうれん草のおひたし	ポテトサラダ
水菜漬け	ロールキャベツ	若竹煮

図 4.1: HTML でデータの収集形式

魚介		
あさりの酒蒸し	さわらの辛味焼き	かきと卵のピリ辛いため
あじの塩焼き	さんまの塩焼き	かきのみそ焼き
塩鮭と野菜のレモン風味いため	あじの和風ハンバーグ	かきフライ
スモークサーモンと玉ねぎのポン酢しょうゆ	いかとえのきのしそいため	かつおのつくりみそダレ和え
いかともやしのからし酢じょうゆ	たいの和風ごまダレサラダ	かれいの煮つけ
いかのトマト煮込み	たこときゅうりの酢の物	さけの和風ムニエル
いかの茎豆和え	たことしめじの辛味しょうゆかけ	さばのおろし煮
いわしのしょうが煮	たこと春雨のキムチ和え	さばの竜田揚げ
まぐろのたたきとトマトモッツァレラのサラダ	たことトマトの地中海風サラダ	さばのみそ煮
うざく	たこのしょうがしょうゆ焼き	ほたて貝の香りバター焼き
エビフライ	たこのにんにく酢みそ和え	ほたて貝のしょうゆ和え
ぶり大根	ツナとポテトのからしマヨネーズ	エビフライ
ぶりの照り焼き	はまちのそぎ造りかわりダレ	えびと春雨のタイ風サラダ

表 4.2: データセットレシピ (肉類)(レシピ番号 40-78)

野菜		
オクラと鶏ささみの梅がつお和え	なすとトマトの冷製サラダ	ほうれん草とえびの鶏いため
ガーリックポテト	なすとひき肉のみそいため	ほうれん草とかきのグラタン
かぶの鶏そぼろあんかけ	なすの揚げ煮	ほうれん草のおひたし
かぼちゃのガーリック風味焼き	なめことえのきのおろし和え	なすトマトとベーコンのチーズ焼き
ポテトサラダ	かぼちゃの煮物	水菜漬け
きゅうりのとうがらし漬け	白菜の刻み漬け	ロールキャベツ
きゅうりもみ	八宝菜	若竹煮
きんぴらごぼう	ピーマンの焼きひたし	新じゃがの煮物
フライパンで作るじゃがいものグラタン風	ゴーやチャンプルー	即席ピクルス
ごぼうと鶏ささ身のサラダ	肉じゃが	大根サラダじゃこドレッシング
小松菜と油揚げの煮ひたし	ベジタブルシチュー	筑前煮
里芋とたこの煮物	回鍋肉片	中華風即席漬け
ししとうがらしとちりめんじゃこのいり煮	さやいんげんのごま和え	ブロッコリーとほたてのチーズ焼き

表 4.3: データセットレシピ (肉類)(レシピ番号 79-117)

豆腐		
揚げ出し豆腐	だし巻き卵	かに玉
厚揚げ・鶏肉・ピーマンのみそいため	豆腐ときのこのキムチ煮	がんもどきと水菜の煮物
あんかけ豆腐	豆腐の野沢菜ちりめんいため	スペイン風オムレツ
マーボー豆腐	半熟卵とレタスのサラダ	

表 4.4: データセットレシピ (肉類)(レシピ番号 118-128)

ご飯		
うなぎ丼	白粥	きのこのリゾット
えびピラフ	たけのこご飯	牛丼
オムライス	ちらしずし	五目炊き込みご飯
親子丼	手巻きずし	五目チャーハン
かきご飯	トマトのプルスケッタ	ハムとチーズのサンドイッチ
ニラ玉雑炊	鶏雑炊	ビーフカレー

表 4.5: データセットレシピ (肉類)(レシピ番号 129-146)

めん		
おかかうどん	ソース焼きうどん	スパゲッティカルボナーラ
かけそば	たらこスパゲッティ	スパゲッティペペコンチーノ
トマトとモッツァレラの冷たいスパゲッティ	カレーうどん	スパゲッティボンゴレ
季節のきのこのクリームスパゲッティ	マカロニサラダ	スパゲッティミートソース
焼きそば	みそラーメン	冷麺

表 4.6: データセットレシピ (肉類)(レシピ番号 147-161)

汁物		
あさりと野菜のトマトスープ	卵スープ	関西風雑煮
いわしのつみれ汁	豆腐となめこのみそ汁	関東風雑煮
かき玉汁	はまぐりの吸い物	けんちん汁
かぼちゃのポタージュ	春雨と野菜のスープ	魚の赤だし
わかめと油揚げのみそ汁	豚汁	レタスとベーコンのソース
わかめともやしのスープ	野菜スープ	さつま汁
しじみのみそ汁		

表 4.7: データセットレシピ (肉類)(レシピ番号 162-180)

その他		
揚げ春巻き	こんにゃくと豚肉の辛味いため	海藻ミックスサラダ梅ドレッシング
アボガドとマグロのバルサミコサラダ	大豆とちりめんじゃこのいり煮	切り干し大根の煮物
ひじきの白和え	天ぷらの盛り合わせ	おでん
ひじきの煮物		

表 4.8: データセットレシピ (肉類)(レシピ番号 181-190)

おやつ		
オートミールのカントリクッキー	サクッとチョコレート	チーズケーキ
オレンジのゼリー	スイートポテト	わらびもち
カスタードプリン	クリームティラミス	

表 4.9: データセットレシピ (肉類)(レシピ番号 701-708)

日付	レシピ番号								健康	便通
2006/11/15	95	95	138	145	186	701	706		1	2
2006/11/16	125	138	138	145	186	701	707		1	2
2006/11/17	118	131	138	138	145	186	701		1	2
2006/11/18	103	138	145	146	701				1	2
2006/11/20	50	83	120	128	138	138	145	701	1	2
2006/11/21	0								1	1
2006/11/22	137	137	145	186	186	701	702	706	1	2
2006/11/23	35	136	138	145	186	701			1	2
2006/11/24	138	145	145	177	183	701	705		1	0

表 4.10: データセットの一部

小松原が開発中の部分は，まだ未完成であるため，本論文の実験では，被験者が直接健康状態を入力したデータを利用することとする． 実験に用いたデータセットは，1ヶ月に食べたレシピ履歴データスクリプトによって生成する．データの中では，日付，食べたレシピ，健康状態，便通の状態を記入されているが，今回はこの中から，日付，食べたレシピ，便通状態の 3 つのデータを切り取って使用した．表 4.10 に示すデータセットの一部の例を使って説明する．1 日に食べたレシピ，健康状態と便通状態を各行に表す．レシピ名の代わりに，あらかじめ付けられたレシピ番号で表す．例えば，表の最初の日に食べたレシピの中のレシピ番号 95 は，対応となるレシピ名は大根サラダじゃこドレッシングである．朝昼晩の順番でなく，レシピ番号の昇順で表す．ここでは，一日中に同じレシピを複数回食べることを認める．便通状態を把握するため，食事アンケートでは便通状態を適当に数値に変換して計算する．例えば，”便が出なかった”，”かたい便がでた”，”やわらかい便がでた”，”水状の便がでた”であれば，それぞれ 0，1，2，3 と変換することをあらかじめ設定しておく．表の便通状態の欄では，対応となる番号を表示される．表 4.10 において，2006/11/19 のデータがチャック漏れや記入の忘れと見なす．2006/11/21 のデータのレシピ番号の欄では，0 を記入されることは何も食べていないことを表す．

4.3 実験の前提条件

便秘とは一般的に，排便が順調に行われない状態のことを言う．しかし，排便の回数には個人差が大きく，便秘をある期間の排便回数で定義することは非常に難しい．今回の実験では， 1 日に排便の回数が 0 であれば，便秘と見なす．一日に複数回に排便した場合，一番悪い状態を入力する．

本論文では，便秘の場合のみ実験を行う．便秘の前日に食べたレシピを相関ルールの前提部とし，翌日に便秘となるのを相関ルールの結論部とする．そして，確信度は便がでなかった日の前日に食べたレシピと，そのレシピの全体の割合とし，サポートは便がでなかった日

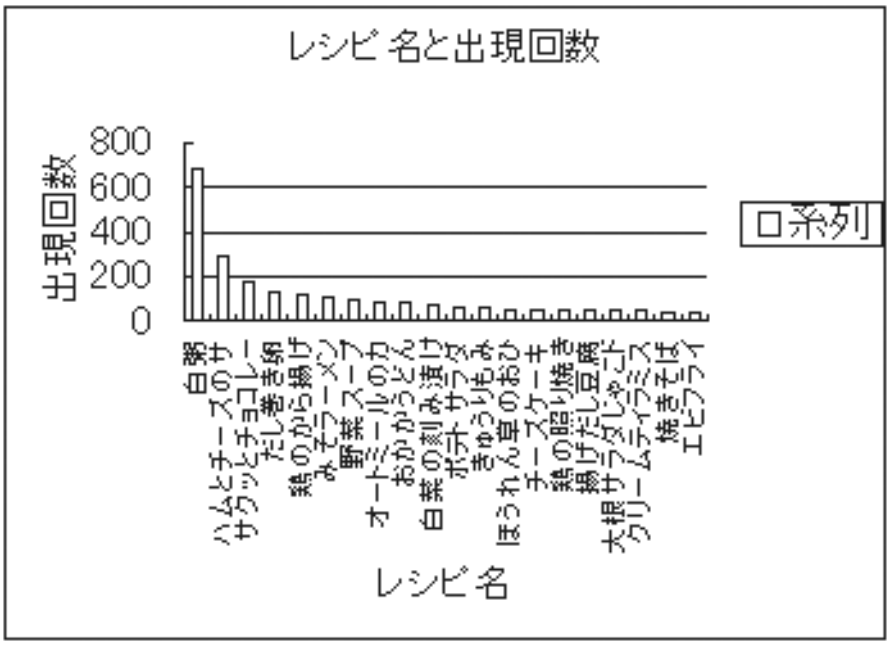

図 4.2: データベースの上位 20 までのレシピ名と出現回数

の前日に食べたレシピは全体の中の割合とする．

第 3 章で述べたように，IBM アルマデン研究所の R.Agrawal によって提案されたアプリオリアルゴリズムでは，候補アイテム集合をつくる時に，その部分集合のすべてが 1 つ前の頻出アイテム集合に出現するもののみを抽出するため，生成必要がない候補アイテム集合が大幅に削減することができるという利点を持つため，今回の実験がアプリオリアルゴリズムの手法を用いて相関ルール抽出し，実験を行う．

本研究では，長さ 1 の頻出アイテム ($< a_1 >, < a_2 > ... < a_{198} >$) が 198 個あるため，長さ 2 の候補アイテム集合 ($< a_1a_1 >, < a_1a_2 > ... < a_1a_{198} >, < a_2a_1 > ... < a_{198}a_{198} >$ と $< (a_1a_2) >< (a_1a_3) > ... < (a_{197}a_{198}) >$) は $198 \times 198 + 198 \times 197 \div 2 = 58707$ 個候補アイテム集合が生成されるという欠点がある．そして，抽出するアイテム集合の長さが長くなればなるほど，必要となる候補アイテムが指数的に増大する．例えば，$minsup = 1$ （すべてのパターンが頻出のとき）が与えられたとき，長さ 198 のアイテム集合をマイニングした場合，長さ 1 の候補アイテム集合は 198 個，長さ 2 の候補アイテム集合は 58707 個…，合計では $2^{198} - 1$ 個となり，記憶容量が多く必要となってしまう．また，データベースのスキャン回数が多くなってしまうため，記憶容量が大きく必要になってしまう．

しかし，198 のレシピには，出現頻度に大きなばらつきがある．データベースにあるレシピ名のそれぞれの出現回数を数えてソートし，出現回数降順で上位 20 までのレシピ名とそれぞれの出現回数を図 4.2 に示す．そのため，最小サポートを適切に設定することで，これらの問題を回避できると考える．

今回，カイ 2 乗検定の有意水準 α が 5%として実験を行う．この場合，相関ルールのカイ 2 乗検定量 $T_{dep} < {x_1}^2(0.05) = 3.841$ であれば，その相関ルールは捨てられる．

第5章　評価手法と結果

本章では，第3章で述べたアプリオリアルゴリズムを用い，第4章で説明した実験環境の元で，提案した食事と健康状態の関連に関する調べる方法を，データについて適用する実験を行う結果について述べる．

本研究では，被調査者に負担を強いずに，再現性・妥当性を確保することに重点を置いている． 1日に食べたレシピの名前を取れば，食事の摂取頻度を詳細に取った時と同様な結果が得られると考え，1日に食べたレシピと健康状態のアンケート調査を行い，食事と健康状態の関連を調べる手順を提案し，調査結果に対してデータマイニングを行う．

蓄積されたデータセットに対して，アプリオリアルゴリズムを用いてデータマイニングを行う．今回，32人の1ヶ月のデータを用い，便秘の前日の1日分のレシピのみからなるデータ集合に着目して単体および組み合わせでデータマイニングを行う．また，便秘の前日だけのデータが必ずしも便秘に影響しているとは限らないと考え，便秘の前の複数日間分のレシピからなるデータ集合に着目して組み合わせてデータマイニングを行う．

まず，便秘の前日の1日分のレシピのみからなるデータ集合に着目して単体の場合について述べる．

本論文では，相関ルールの前提部 X をレシピAとし，相関ルールの結論部 Y を次の日に便秘になることとする．最小確信度を50%とし，カイ2乗分布の有意水準を5%とする．また，1日のデータを一つのトランザクションとする．データセットにおいて，各人に対してそれぞれの便通状態が0であれば，対応となる前日のレシピを取り出す．各人のデータの中では，連続していないデータがある（例えば，記入漏れや何も食べていない日など）．このため，便秘の前日が存在しない場合は無視する．便秘の前日に食べた各単体レシピの回数と，それぞれのレシピが全体の中の回数を数える．便秘の前日に食べたそれぞれのレシピが全体の割合即ち確信度を求め，最小確信度を満たす相関ルールを抽出する．さらに，抽出された相関ルールに対して，自由度1のカイ2乗検定を行う．各人のデータに対してそれぞれの便秘の日数とデータ総数を数え，抽出された相関ルールの前提部に対し，それぞれのサポート S_X，S_Y，S_{XY} を求め，式3.1に代入し，それぞれの相関ルールのカイ2乗検定量 T_{dep} を求める．求められたカイ2乗検定量の値は（カイ2乗分布表より有意水準5%の時，検定量 T_{dep} の値が3.841となる）3.841より大きい相関ルールの前提部と結論部の相関が強いため，価値のある相関ルールとして出力される．実験結果は表5.1に示す．

今回，全ての人に使えるものを探すのが目的なので，この場合個人のデータに対して，便秘の前日の1日分のレシピのみからなるデータ集合に着目して単体データマイニングを行

レシピ	カイ 2 乗検定量	確信度	サポート
酢豚	6.849704	0.833333	0.006361
あさりの酒蒸し	4.033251	1	0.002545
関西風雑煮	4.033251	1	0.002545
大豆とちりめんじゃこのいり煮	7.856304	0.727273	0.010178

表 5.1: レシピ単体の実験結果

レシピの組み合わせ	カイ 2 乗検定量
ハムとチーズのサンドイッチ，サクッとチョコレート	3.917093

表 5.2: 前日のみのレシピの組み合わせの実験結果

わず，全体の場合のみの実験を行う．

便秘の前日の 1 日分のレシピのみからなるデータ集合に着目して組み合わせの場合について述べる．

今回のデータセットは 32 人分の 1ヶ月のデータを用い，レシピが全部で 198 個あるため，同じレシピの出現確率が非常に少ないので，最小確信度設定せずに，最小サポートを 20 とし，カイ 2 乗分布の有意水準を 5%とする．1 日のデータを一つのトランザクションとする．この場合，相関ルールの前提部 X を便秘の前日に食べたレシピの組み合わせとし，次の日に便秘となることを結論部 Y とする．便秘の前日のレシピをアプリオリアルゴリズムによって，最小サポートを満たす頻出レシピの組み合わせの集合を生成させる．同じレシピの組み合わせに対して，全体の頻出レシピの組み合わせの集合も生成する．それぞれのレシピの組み合わせのサポートをカウントする．さらに，最小サポートを満たす相関ルールに対して，自由度 1 のカイ 2 乗検定を行う．各人のデータに対してそれぞれの便秘の日数とデータ総数を数え，抽出された相関ルールの前提部に対し，それぞれのサポート S_X，S_Y，S_{XY} を求め，式 3.1 に代入し，それぞれの相関ルールのカイ 2 乗検定量 T_{dep} を求める．検定量 T_{dep} の値は 3.841 を満たす相関ルールのみ出力される．実験結果は表 5.2 に示す．

同様に，二日前，三日前，四日前の実験結果がそれぞれ表 5.3，5.4，5.5 に示す．

次に，便秘の前の複数日間分のレシピからなるデータ集合に着目して組み合わせてデータマイニングを行う実験を述べる．ここでは，便秘の前の複数日間分を N とする．$2 \leq N \leq 4$ について実験を行う．

まず，便秘の前二日間分のデータを一つのトランザクションとする場合について述べる．この場合では，相関ルールの前提部 X を便秘の前二日間分のレシピの組み合わせとし，二日後に便秘となることを結論部 Y とする．便秘の前二日間食べたレシピを取り出し，アプリオリアルゴリズムによって，便秘の前の二日間分のレシピの組み合わせの頻出アイテム集合を生成される．同じレシピの組み合わせに対して，全体の頻出レシピの組み合わせの集合も生成される．それぞれのレシピの組み合わせのサポートをカウントする．さらに，最小サ

レシピの組み合わせ	カイ 2 乗検定量
だし巻き卵，白粥	8.15726
白粥，ハムとチーズのサンドイッチ	5.02715
ハムとチーズのサンドイッチ，サクッとチョコレート	10.866193

表 5.3: 二日前のみのレシピの組み合わせの実験結果

レシピの組み合わせ	カイ 2 乗検定量
鶏のから揚げ，白粥	5.392515
だし巻き卵，白粥	11.206815
白粥，サクッとチョコレート	3.889798
ハムとチーズのサンドイッチ，サクッとチョコレート	20.681059

表 5.4: 三日前のみのレシピの組み合わせの実験結果

ポートに満たす相関ルールを自由度 1 のカイ 2 乗検定を行う．各人のデータに対してそれぞれの便秘の日数とデータ総数を数え，抽出された相関ルールの前提部に対し，それぞれのサポート S_X，S_Y，S_{XY} を求め，式 3.1 に代入し，それぞれの相関ルールのカイ 2 乗検定量 T_{dep} を求める．検定量 T_{dep} の値は 3.841 を満たす相関ルールのみ出力される．実験結果は表 5.6 に示す． 同様に，前三日間分と前四日間分のレシピの組み合わせの実験結果が得られる．表 5.7 と表 5.8 に示す．

レシピの組み合わせ	カイ 2 乗検定量
だし巻き卵，白粥	11.222124
ハムとチーズのサンドイッチ，サクッとチョコレート	19.64843
白粥，ハムとチーズのサンドイッチ，サクッとチョコレート	4.101117

表 5.5: 四日前のみのレシピの組み合わせの実験結果

レシピの組み合わせ	カイ 2 乗検定量
ハムとチーズのサンドイッチ，サクッとチョコレート	4.876865

表 5.6: N=2 の場合のレシピの組み合わせの実験結果

レシピの組み合わせ	カイ 2 乗検定量
白粥，ハムとチーズのサンドイッチ	4.3198314
白粥，おかかうどん	3.900086
白粥，サクッとチョコレート	4.993071
ハムとチーズのサンドイッチ，サクッとチョコレート	6.900366

表 5.7: N=3 の場合のレシピの組み合わせの実験結果

レシピの組み合わせ	カイ 2 乗検定量
白粥，ハムとチーズのサンドイッチ,	5.176787
ハムとチーズのサンドイッチ，サクッとチョコレート	4.040732

表 5.8: N=4 の場合のレシピの組み合わせの実験結果

第6章　考察

アプリオリアルゴリズムによって，食事と健康状態に関するデータに対してデータマイニングを行う実験では，便秘の前日の１日分のレシピのみからなるデータ集合に着目して単体および組み合わせでデータマイニングを行うことと，便秘の前の複数日間分のレシピからなるデータ集合に着目して組み合わせに着目する．

まず，レシピの単体の場合について考察する．酢豚，大豆とちりめんじゃこのいり煮，あさりの酒蒸しと関西風雑煮を食べると，次の日に便秘になる可能性が高いと予測することができる．

次に，レシピの組み合わせの場合について述べる．まず，便秘の前日の１日分のレシピのみからなるデータ集合に着目して組み合わせでデータマイニングを行う実験についての考察を述べる．便秘の前日のみに着目すると，ハムとチーズのサンドイッチとサクッとチョコレートの組み合わせを食べると，次の日に便秘になる可能性が高いといえる．便秘の二日前の場合では，出し巻き卵と白粥，白粥とハムとチーズのサンドイッチ，そして，ハムとチーズのサンドイッチとサクッとチョコレートの三つの組み合わせのどれかを食べると，二日後に便秘になる可能性が高いと予想することができる．そして，便秘の前の二日間分のレシピからなるデータ集合に着目し，組み合わせてデータマイニングを行う実験結果と合わせて考察すると，便秘の前日と二日前共にハムとチーズのサンドイッチとサクッとチョコレートの組み合わせという結果になり，便秘の前の二日間の結果に影響していると考えられる．次に，便秘の三日前の実験結果と便秘の前の三日間分の実験結果から見ると，白粥，ハムとチーズのサンドイッチとサクッとチョコレートの三つの組み合わせのうちのどれかを食べると，三日後に便秘になる可能性が高いと考えられる．最後に，便秘の四日前の実験結果が白粥，ハムとチーズのサンドイッチとサクッとチョコレートの組み合わせの他に，出し巻き卵や鶏のから揚げの出現回数が高いにも関わらず，便秘の前の四日間分の実験結果では，白粥，ハムとチーズのサンドイッチとサクッとチョコレートの組み合わせが食べると，四日後に便秘になる可能性が高いという結果から，便秘の前の四日間分をまとめて見ることによって，四日後に便秘になることに関する予測が不十分であることがわかる．以上より，ハムとチーズのサンドイッチとサクッとチョコレートの組み合わせを食べると，便秘になる可能性が高いと予想することがで来るであろう．また，便秘の前日だけのデータが必ずしも便秘に影響しているとは限らないと言える．

レシピの単体の場合とレシピの組み合わせの場合では，全然違う結果が得られた．これは，レシピの単体の場合で得られた実験結果のレシピの組み合わせは，全体の中の出現回数，即

ちサポートが少なく，設定された最小サポート 20 以下であるため，アプリオリアルゴリズムによって枝刈りされたためだと考えられる．

今回の実験で使用されたレシピの項目に偏りがある．例えば，果物，ヨーグルトなどいわゆる便秘の改善に良い食べ物と，サプリメントや薬など便秘に影響するレシピがないことなど，限られたレシピで実験を行ったため，調査として十分とは言えないところがある．そして，一日に数回排便がした場合の考慮をしていないことと，便の固さの判断の個人差も考慮せずに実験を行ったため，得られた結果も多少のノイズを含んでいると考えられる．また，食べ物の順序を考慮したマイニングは行っていないこと，データ数が不足していること，レシピの数が多いなども結果に多く影響する原因と考えられる．また，便秘に影響とする生理も一つの要素として挙げられる．従って，実験で得られた結果はあくまでも予測としか言えないと考える．

第7章　まとめ

本論文では，データマイニングを用いて効果的に健康によい食べ物，健康によくない食べ物の特徴を発見することで，食べ物の選択および健康状態の管理に役立つ指標を見出すことを目的とし，データマイニングを用いて食事と健康状態の関連の調べに関する手順を提案し，実験を行った．今回の実験では，32 人の 1ヶ月のデータを用い，便秘の前日の 1 日分のレシピのみからなるデータ集合に着目して単体および組み合わせでデータマイニングを行ったことと，便秘の前日だけのデータが必ずしも便秘に影響しているとは限らないと考え，便秘の前の複数日間分のレシピからなるデータ集合に着目し，組み合わせてデータマイニングを行った．その結果，レシピの単体の場合では，酢豚，アサリの酒蒸し，関西風雑煮と大豆とちりめんじゃこのいり煮を食べると，次に日に便秘になる可能性が高いと予想することができる．また，レシピの組み合わせの場合では，ハムとチーズのサンドイッチとサクッとチョコレートの組み合わせを食べると，次の日，二日後，三日後と四日後に便秘になる可能性が大きいと予想することができた．

今後，より確信度高い予想を得られるため，データマイニングの対象とするデータベースの形式をレシピより，素材まで分類する必要があると考えられる．また，長期的に渡ってデータを蓄積することになり，アプリオリアルゴリズムの記憶量が大きく必要となるため，新たな手法を用いてマイニングする必要があると考えられる．今回は，ホームページからデータ入力を行い，コンピュータ上でデータマイニングを行っている．今後は，データ入力とデータマイニングを行う i アプリを開発したいと考えている．

謝辞

指導教官である城和貴教授には，本研究だけでなく日本での留学生活においても，暖かい御指導と多大な助言を頂き，本当に大変お世話になりました． 心から深く感謝しております．この場を借りて，心から厚くお礼を申し上げます．

また，高田雅美先生には，本研究を行うことにあたり，丁寧な御指導を頂き，研究生活においてもいろいろと大変お世話になりました．心から感謝しております．どうもありがとうございました．

最後に，城研究室の皆様には，本研究のためのたくさんの意見を頂くだけではなく，食事のアンケートの収集に御協力を頂くなど様々な場面で大変お世話になりました．これまで充実した楽しい研究生活を過ごせましたのも皆様のおかげです．本当にありがとうございました．

参考文献

[1] 食事摂取プランの研究：http://www.v350f200.com/kanri/kankei_2g.html

[2] 日本人の食事摂取基準：http://www.mhlw.go.jp/houdou/2004/11/h1122-2.html

[3] 佐々木敏”健康的な食生活習慣形成を目指した食事摂取基準”

[4] 食事調査法の開発：http://www.nih.go.jp/eiken/main_adult.html

[5] 国立健康・栄養研究所：http://www.nih.go.jp/eiken/programs/ekigaku_shokuji.html

[6] 食物摂取頻度調査：http://www2.eiyo.shikoku-u.ac.jp/eiyokun/soft/FFQg/FFQg.htm

[7] 竹内裕之，児玉直樹，橋口猛志，林同文”個人健康管理を目的とした健康データマイニングシステム”，DEWS2006 論文集，1B-ill，2006

[8] Ian H.Witten, Eibe Frank, ”Data Mining - Practical Machine Learning Tools and Techniques with Java Implementations” Morgan Kaufmann Publishers, 1999

[9] W.Frawley and G.Piatetsky-Shapiro and C.Matheus, Knowledge Discovery in Databases: An Overview. Al Magazine, 213-228, Fall 1992

[10] 福田剛志，森本康彦，徳山豪”データマイニング”共立出版，2001

[11] R.Agrawal, A.Arning, T.Bollinger, M.Mehta, J.Shafer, and R.Srikant, The Quest data mining system. In Proceedings of the International Conference on Knowledge Discovery and Data Mining, 1996

[12] 小松俊介，山名早人”アイテム間の距離を考慮した Sequential Pattern Mining の提案”，DE2005-72，pp.43-48，2005

[13] 決定木の定義：http://www.engr.ie.u-ryukyu.ac.jp/ taka/soturon/genkou/node21.html

[14] 黄嵩”強化学習と決定木学習による汎用エージェント構成の試み”，2004

[15] L.Hyafil, R.Rivest. Constructing optimal binary decision tree is NP-complete. Information Processing Letters, 5:15-17, 1976

[16] Quinlan, J. R. Induction of decision trees. Machine Learning, 1:81-106, 1986

[17] Quinlan, J. R. C4.5:Programs for Machine Learning. Morgan Kaufmann, 1993

[18] 決定木の利点と欠点:http://mikilab.doshisha.ac.jp/dia/research/report/2005/0712/001/report20050712

[19] 神嶌敏弘, ”データマイニング分野のクラスタリング手法 (1)—クラスタリングを使ってみよう！—”, 人工知能学会誌, vol.18, no.1, pp.59-65, 2003

[20] クラスタリングについて：http://www.kamishima.net/jp/clustering/