個人健康管理システムのための自動相関ルール抽出アルゴリズム

Automated Rule Induction Algorithm for a Personal Dynamic Healthcare System

竹内 裕之[1)] 児玉 直樹[2)]
橋口 猛志[3)] 林 同文[3)]

Hiroshi TAKEUCHI Naoki KODAMA
Takeshi HASHIGUCHI Doubun HAYASHI

本論文で報告する個人健康管理のための自動ルール抽出アルゴリズムは、携帯電話と web テクノロジーを活用した個人健康管理システム上に時系列的に蓄積された生活習慣と健康状態に関する日常のデータからシステムユーザ個人にとって有用な生活習慣と健康状態の関係を抽出する．この個人健康管理システムは、ユーザが携帯電話を用いて入力した日常のデータをインターネット経由で web アプリケーションサーバに蓄積しサーバが生活習慣と健康状態に関する解析結果をユーザの携帯電話に返す仕組である．あるボランティアユーザの蓄積されたデータを基にここで開発したアルゴリズムを用いてデータマイニングを行った結果、生活習慣と日常の血圧の間に有用な関係を見出した．

The automated-rule-induction algorithm reported here extracts personally useful information concerning lifestyles and health conditions from daily time-series personal health and lifestyle data stored on a personal dynamic healthcare system by using mobile phone and web technologies. This system enables users to input their daily data through a mobile phone and to transfer these data to a web-application server via the Internet. The web application server provides a data-mining service and uses mobile phones to inform users of important rules concerning their lifestyles and health conditions. Healthcare-data mining of the stored time-series data of a volunteer user based on the automated-rule-induction algorithm generated some useful rules concerning their lifestyles with blood pressures.

1. はじめに

最近リモートの患者をケアする遠隔医療へのWebテクノロジーの応用が世界的に注目を集めている[1-3]．急速に発展したWebテクノロジーは、さらに予防医学や健康管理の面でもその活用が期待されている．特に、わが国では少子高齢化の進展が著しく、国民一人一人が健康で長生きして若年層の負担を軽減することが喫緊の課題となっており、病気の一次予防や健康増進のために日常の個人ベースの健康管理が必要であることが指摘されている[4]．

そこで、我々は携帯電話とWebテクノロジーを活用した個人健康管理システムを開発した[5],[6]．このシステムは、携帯電話を端末として日常の生活習慣データと健康データをインターネット経由で時系列的にサーバコンピュータに蓄積する仕組である．蓄積されたデータはその統計を判りやすいグラフ表示で見ることができる他、個人の生活習慣と健康状態の間に何らかの規則性が見出せれば相関ルールとしてユーザの携帯電話に通知する．そして、これらの情報を参考にユーザが自分で自分の健康管理を行うことを期待している．本論文では、開発した個人健康管理システムにおいて、時系列的に蓄積されたデータから生活習慣と健康状態の相関ルールを自動的に抽出する手法について述べる．

2. 健康データマイニングのコンセプト

健康データマイニングのコンセプトを図１に示す。ここでは、個人の現在の健康状態がなんらかの形で日常の生活習慣の影響を受けていると仮定する．そして、その関係は複数の項目が絡んだ複雑なもので、個人差も大きいものと考える．健康データマイニングの目的は、日常の生活習慣データと健康データを個人毎に時系列的に蓄積し、その中から生活習慣と健康状態の間になんらかの規則性を見出し個人毎のルールとして抽出することである[7]．

したがって、健康データマイニングでは、生活習慣データ項目を入力変数（独立変数）、健康データ項目を出力変数（ターゲット変数）として位置付け、「生活習慣データ $Y=y$ ならば健康データ $X=x$ の傾向がある」といったルールを個人毎に抽出する．即ちルールの前提部には生活習慣データ項目が、結論部には健康データ項目が含まれる．システムのユーザはこのような個人毎のルールを自己の健康管理や健康増進のために役立てることができる．

なおここではルールをシンプルにするために生活習慣データ項目間の関連性については考慮しなかった．

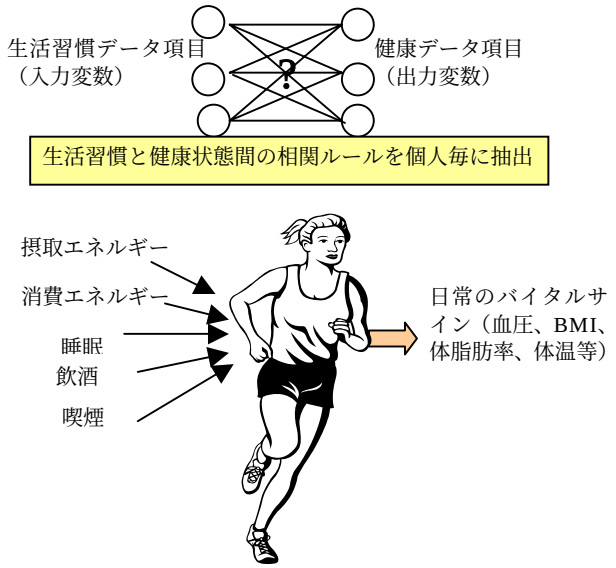

図１ 健康データマイニングのコンセプト[7]
Fig.1 Concept of Healthcare Data Mining [7]

[1)] 正会員 高崎健康福祉大学健康福祉学部医療福祉情報学科 htakeuchi@takasaki-u.ac.jp
[2)] 非会員 高崎健康福祉大学健康福祉学部医療福祉情報学科 kodama@takasaki-u.ac.jp
[3)] 非会員 東京大学大学院医学系研究科健康医科学創造講座 {hashiguchi-mi,hayashi-tky}@umin.ac.jp

3. 自動ルール抽出のアルゴリズム

3.1 ルールの生成方法

開発した個人健康管理システムでは、インターネット経由で不特定多数のユーザが日常のデータを蓄積し、それに対してタイミングよくデータ解析（健康データマイニング）を行い有効なルールを提示する必要がある．従って、データがある程度蓄積されると自動的に解析が実行されることがサービスとして必須である．そこで我々は、多くのデータセットの中から自動的にルールを抽出するのに適した手法として、情報理論に基づくアソシエーションルール解析手法であるITRULEアルゴリズム[8]を用いることにした．

ITRULEアルゴリズムは、

$$\text{If } Y=y, \text{ then } X=x \text{ with probability } p \qquad (1)$$

という簡単なルールを生成する．健康データマイニングでは、Yは生活習慣データ項目でyはその条件、Xは健康データ項目でxはその値である．結論部は関心ある健康データ項目１つに制限するが、前提部を複数にすることを許容する．即ち、Zを別の生活習慣データ項目、zをその条件として、

$$\text{If } Y=y \text{ and } Z=z, \text{ then } X=x \text{ with probability } p \qquad (2)$$

なるルールを許容する．ルール(2)をルール(1)の特殊化と呼ぶ．

ITRULEアルゴリズムでは、蓄積された多くのデータセットから有効なルールを生成するメカニズムとして式(3)に示すJ測度[8]を用いる．

$$J(x|y)=p(y)\left(p(x|y)\log\frac{p(x|y)}{p(x)}+(1-p(x|y))\log\frac{(1-p(x|y))}{(1-p(x))}\right) \qquad (3)$$

式(3)において、大括弧内は$Y=y$という事象が起きた場合にXの値に関して得られる情報の大きさを表す．ルールの場合には$X=x$か$X=\overline{x}$なので２項の和になっている．即ち大括弧内は$Y=y$という前提がある場合とない場合でXの値に関する確率分布がいかに異なるかという尺度であるとも言える．J測度はこれに母集団において$Y=y$という事象が起きる確率$p(y)$を掛けたもので、この値が大きいルールが良いルールということになる[8]．

実際にはルールのJ測度は、蓄積されたユーザ毎の時系列データセットをサンプル集団として、$p(y)$をルールの前提条件がデータセットからのサンプルと一致する確率、$p(x)$をルールの結論がデータセットからのサンプルと一致する確率、$p(x|y)$を前提条件で条件付けられたルールの結論の条件確率、として計算する．

具体的なルールの自動生成プロセスは以下に述べるように、ユーザ毎の時系列データセットをサンプル集団としJ測度が大きいルールが生き残るように実行される．

(a) 各出力フィールドX_i（関心のある健康データ項目）を順番に処理する．即ち、次の出力フィールドを対象とする前に、現在の出力フィールドに対してすべてのルールを生成する．

(b) 各出力フィールドX_iに対して、ある１つの値x_kを選択する．そして、次の値を対象とする前に、現在の値を結論とする全てのルールを生成する．

(c) 各値x_kに対して、それぞれの入力フィールドY_jを選択する．

(d) 各入力フィールドY_jに対して、それぞれの条件y_qを選択する．それぞれの条件は、入力フィールドのデータ型によって異なる．

（ｉ）シンボル値フィールドの場合、フィールドの各値が条件となる．

（ii）数値型フィールドの場合、値はソートされ、各値が２分割境界としてテストされる．具体的には、各分割に対してJ測度が算出され、最も大きいJ測度を持つ分割が選択される．従って、「選択された分割値より大きい」と「選択された分割値以下」の２つの条件のみが可能となる．

(e) ルール[$Y_j=y_q$ならば$X_i=x_k$]に対して、J測度を算出する．

(f) 算出されたJ測度の値が、ルール格納テーブル中の同じ結論（$X_i=x_k$）を持つルールの中で最大のJ値より大きい場合、またはテーブル中のルール数が設定された最大数未満でかつ設定された最小サポート率および最小確信度基準を満たしていればテーブルにルールが格納され（必要に応じてJ値の低いルールが置き換えら）、さらに式(4)で示すJ_s値が評価される．それ以外の場合は、次の入力フィールドの条件に進む．ここで、サポート率とはサンプルデータセット中で、ルールの前提部が真のレコード数の割合、確信度とはルールの前提部が真のレコード中で、結論部が真のレコード数の割合である．

(g) J_sの値が、ルール格納テーブル中のルールのJ値の最小値より大きい場合にはルール特殊化（前提条件の追加）を試みる．

(h) すべての入力フィールド条件、入力フィールド、出力フィールド値、および出力フィールドを検討し終わるまで処理を繰り返す．

ここで、前述のJ_s値とルールの特殊化について説明する．ITRULEアルゴリズムではテーブルにルールが格納されると、そのルールを特殊化する（前提条件にさらに別の条件を追加する）潜在的な利点があるかどうか調べる．このために式(4)で定義されるJ_s値を評価する[8]．

$$J_s=\max\left(p(y)p(x|y)\log\left(\frac{1}{p(x)}\right),\ p(y)(1-p(x|y)\log\left(\frac{1}{1-p(x)}\right)\right) \qquad (4)$$

J_s値は特殊化を行った場合のJ値の上限値である．従って、この値がその時点でのルール格納テーブル中で同じ結論部を持つルールにおける最小のJ値より大きい場合にはルールの特殊化を試みる．即ち、前提条件にさらに別の入力フィールドを追加して、元の特殊化されていないルールの場合と同様な処理（(a)～(h)）を実行する．そして、特殊化されたルールのJ値がその時点でのルール格納テーブル中の最小のJ値を超えていれば、そのルールを置き換える．

特殊化は、追加する入力フィールドがなくなるまで処理を続けることが出来るが、健康データマイニングでは、前提条件が多いあまりにも複雑なルールは適当ではないので、２次のオーダーで止める（即ち前提条件は最大２つとする）ことにした．

3.2 入力変数定義

本研究の最終目標は、インターネット上の不特定多数ユーザの大量の日常データを処理することなので、あまりにも多くの（有効でない）入力変数はいたずらにコンピュータの処理時間を消費するだけである．

そこで、関心のある健康データ項目（出力変数）と入力変数になりうる生活習慣データ項目の時系列データをもとに、相関を事前にチェックし入力変数（フィールド）を自動定義

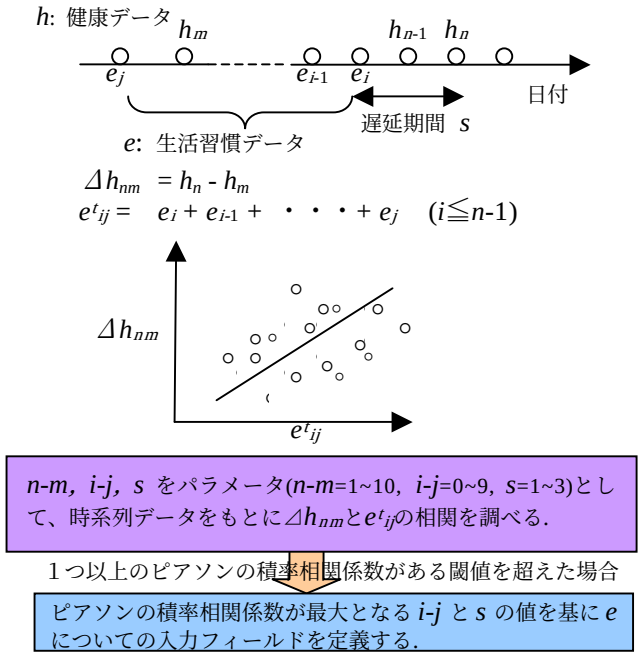

図2　入力フィールド自動定義のアルゴリズム

Fig.2 Algorithm for Defining Input Field Automatically

する手法を開発した．以下、図2を用い入力変数（フィールド）の自動定義のアルゴリズムについて説明する[7]．図2において、hは関心のある健康データ項目を、eは入力変数（フィールド）の候補となる生活習慣データ項目を表す．h_nはhのn日におけるデータ値、e_iはeのi日におけるデータ値、sは遅延期間である．ここで、

$$\Delta h_{nm} = h_n - h_m \qquad (5)$$

なる量と、

$$e^t_{ij} = e_i + e_{i-1} + \cdots + e_j \qquad (6)$$

なる量を定義する．遅延期間は $s = n - i$ で定義する．

次に蓄積された時系列データについて、n-m, i-j, sをパラメータとして変化させながら(n-m=1~10, i-j=0~9, s=1~3)、Δh_{nm}とe^t_{ij}の間のピアソンの積率相関係数を計算する．ここで、n日より前の生活習慣データがn日の健康データに影響を与えると仮定する($i \leqq n$-1)．各(n-m, i-j, s)のセットにつきピアソンの積率相関係数を計算し、もし、1つ以上の相関係数がある閾値R_sより大きいものがあれば、そのeをhに対する入力変数として採用する．そして、実際の入力フィールドは相関係数が最大となる(n-m, i-j, s)のセット$((n\text{-}m)_{max}, (i\text{-}j)_{max}, s_{max})$をもとに自動定義する。例えば、$(i\text{-}j)_{max}$=2, s_{max}=2 であれば、eに関わる入力フィールドを

$$e_i + e_{i-1} + e_{i-2}\ (i=n\text{-}2) \qquad (7)$$

と自動定義する．$(i\text{-}j)_{max}$が大きいということは、長期間の生活習慣データの蓄積が現在の健康データに影響を与え、s_{max}が大きいということは、生活習慣データが遅れをもって現在の健康データに影響を与えるということになる．

生活習慣データ項目eの値が数値でなくシンボル値の場合は、時系列データに基づくhとの相関係数は、シンボル値を適当に数値に変換して計算する．例えば、シンボル値が“多い”、“普通”、“少ない”、であれば、それぞれ3, 2, 1 と変換する．例として、e_i=”多い“、e_{i-1}=”少ない“、であれば、$e_i + e_{i-1} = 3 + 1 = 4$ とする．ただし、入力変数として採用されルールマイニングを行うときには入力フィールドの値はシンボル値をそのまま用いる．

3.3 出力変数定義

健康データマイニングにおける出力変数（フィールド）は、関心ある健康データ項目である．健康データ項目hの値がシンボル値の場合は、Δh_{nm}を計算するときには前項で述べたシンボル値を持つ生活習慣データ項目と同じように数値に変換する．ただし、ルール生成プロセスでは出力変数（フィールド）値はシンボル値をそのまま用いる．

一方、健康データ項目 h の値が数値の場合は、全ての時系列データを“高い”、“中間”、“低い”というシンボル値を持つ3つのクラスに分類する．この分類において3つのクラスの境界値は、それぞれのクラスのデータ頻度が同程度になるように自動設定する．

4. 健康データマイニングの実例

4.1 ボランティアユーザの蓄積データ

あるボランティアユーザが蓄積した日常の生活習慣データ項目と健康データ項目およびそれぞれのデータ型を表1に示す．データはほぼ毎日の入力で約1年間蓄積された．

表1 ボランティアユーザが蓄積したデータ項目とデータ型

Table 1 Registered Data Items and Their Data Types of a Volunteer User

生活習慣データ項目（データ型）	健康データ項目（データ型）
運動による消費エネルギー（数値）	最大血圧（数値）
食事による摂取エネルギー（数値）	最小血圧（数値）
アルコール摂取量（シンボル値）	脈拍数（数値）
睡眠時間（数値）	
睡眠の深さ（シンボル値）	
ストレス（シンボル値）	

運動による消費エネルギー（kcal）はユーザが身に付けている歩数計とフィットネスクラブの記録から、食事による摂取エネルギー（kcal）は朝、昼、夜のメニューからユーザの判断でデータ登録した．

アルコール摂取量、睡眠の深さ、ストレスはユーザの判断でそれぞれ5段階（飲み過ぎ、多い、適度、少ない、非常に少ない）、3段階（ぐっすり、やや浅い、あまり眠れなかった）、3段階（多い、普通、少ない）のシンボル値でデータ登録した．

血圧（mmHg）と脈拍（回/分）は、オシロメトリック法による自動血圧計を用い、家庭でほぼ毎日朝同じ条件で測定している．血圧は測定のエラーを少なくするために、3回計測しその平均をデータ登録した．

4.2 自動定義された入力変数（フィールド）

ボランティアユーザの関心事は、日常の血圧に与える生活習慣の影響にあった．従って、出力変数は最大血圧（心臓収縮期血圧）と最小血圧（心臓拡張期血圧）である．そこで、表1の生活習慣データ項目とこれらの健康データ項目の時系列データを基に3.2節で述べた手法により入力変数（フィールド）を定義した．この時、ピアソンの積率相関係数はそ

れぞれのデータ項目について時系列の最初の 80 データを基に計算し、相関が有意であると判断できる相関係数の基準を 0.3 と仮定してR_s=0.3 と設定した．最小血圧についての結果を表２にまとめた．

表２　最小血圧に対して自動定義された入力変数（フィールド）
Table 2 Automatically Defined Input Fields for Diastolic Blood Pressure

入力変数	r	定義された入力フィールド
消費エネルギー	-0.333	e_i+e_{i-1}+・・・+e_{i-5}(i=n-1)
摂取エネルギー	0.337	e_i(i=n-2)
アルコール摂取量	0.377	e_i(i=n-2)
ストレス	0.330	e_i(i=n-1)
実効睡眠時間	-0.312	e_i+e_{i-1}+・・・+e_{i-6}(i=n-1)

表２において、rは 3.2 節で述べたパラメータセット($(n\text{-}m)_{max}$, $(i\text{-}j)_{max}$, s_{max})におけるピアソンの積率相関係数の値で、このパラメータセットに基づき入力フィールドは定義されている．表２における実効睡眠時間は睡眠時間を睡眠の深さの値（1（ぐっすり)、2（やや浅い)、3（あまり眠れなかった))の平方根で割って重み付けをしたものである．なお重み付けの方法はこれに限ったことではない．

4.3 生成されたルール例

S_s=0.04、C_s=0.65 と設定して自動生成された最小血圧に関する８個のルールを図３にまとめた．ここで、インスタンスIはルールの前提が真のレコード数、サポート率SはIを総レコード数で割った値、確信度Cはルール全体が真のレコード数をIで割った値である．ルールは$S\times C$の値が大きい順にソートして表示している．ここでルール４はルール３の特殊化で確信度がさらに高くなっている．

２日前のアルコール摂取量と摂取エネルギーおよび昨日のストレスが朝の血圧に大きな影響を与えていることが判る．これらのルールは個人毎に異なると考えられ、ユーザはこのような個人特有の情報に基づいて個々の生活習慣の改善を図ることができる．

出力フィールド: 最小血圧
レコード総数: 333
高い: ≧ 86 mmHg　　I : インスタンス
中間: 82-85 mmHg　　S:サポート率
低い: ≦ 81 mmHg　　C: 確信度
ルール導出条件: 最大前提数 = 2, S≧0.04, C≧0.65

	前提部	結論部	I	S	C
1	[ストレス1=少ない, 摂取1< 1625 kcal]	[最小血圧=低い]	40	0.12	0.7
2	[酒量1=飲み過ぎ]	[最小血圧=高い]	36	0.108	0.72
3	[ストレス1=多い]	[最小血圧=高い]	35	0.105	0.69
4	[ストレス1=多い, 消費6< 1819 kcal]	[最小血圧=高い]	27	0.081	0.81
5	[酒量1=飲み過ぎ, 摂取1>2025 kcal]	[最小血圧=高い]	22	0.066	0.86
6	[ストレス1=少ない, 酒量1=飲み過ぎ]	[最小血圧=高い]	23	0.069	0.78
7	[ストレス1=普通, 睡眠7> 41.67 時間]	[最小血圧=中間]	16	0.048	0.75
8	[酒量1=少ない, 摂取1> 1945 kcal]	[最小血圧=中間]	16	0.048	0.69

・ストレス1: 昨日のストレス
・摂取1: 2日前の摂取エネルギー
・酒量1: 2日前のアルコール摂取量
・消費6: 昨日から6日間の消費エネルギー
・睡眠7: 昨日から7日間の実効睡眠時間

図３ 最小血圧に関して自動生成されたルール
Fig.3 Automatically Generated Rules for Diastolic Blood Pressure

5. まとめ

開発した個人健康管理のための自動ルール抽出アルゴリズムを用いて個人健康管理システムのボランティアユーザの生活習慣と日常の血圧の関係において有用な情報を抽出できることを示した．今後、多くのボランティアユーザを募り、システムの有用性を評価し普及を図る．

[文献]

[1] N. H. Lovell, F. Magrabi, B. G. Celler, K. Huynh, and H. Garsden, “Web-based acquisition, storage, and retrieval of biomedical signals,” IEEE Eng. Medicine and Biology, vol.20. no.3, pp.38-44, 2001.

[2] J. Cai, S. Johnson, and G. Hripcsak, “Generic data modeling for home telemonitoring of chronically ill patients,” Proc. AMIA Symp. pp.116-120, 2000.

[3] C. Mazzi, P. Ganguly, and M. Kidd, “Healthcare application based on software agents,” Medinfo 2001 Proceedings, pp.136-140, 2001.

[4] T. Hashiguchi, H. Takeuchi, and A. Uemura, “Highly advanced healthcare support services for the 21st century,” Hitachi Review, vol.50, no.1, pp.2-7. 2001.

[5] 竹内裕之, 橋口猛志, 新谷隆彦, “日常の健康管理を目的とした個人対応動的データベース” 医療情報学 vol.23, no.6, pp.497-502, 2004.

[6] H. Takeuchi, T. Hashiguchi, and T. Shintani, “Personal dynamic healthcare system utilizing mobile phone and web technologies,” Proc. 2nd Int. Conf. on Advances in Biomedical Signal and Information Processing, pp.304-307, 2004.

[7] H. Takeuchi, N. Kodama, T. Hashiguchi, and N Mitsui,” Healthcare data mining based on a personal dynamic healthcare system,” Proc. 2nd Int. Conf. on Computational Intelligence in Medicine and Healthcare, pp.37-43, 2005.

[8] P. Smyth and R. M. Goodman, “An information theoretic approach to rule induction from databases,” IEEE Trans. Knowledge and Data Engineering, vol.4, no.4, pp.301-316, 1992.

竹内 裕之　Hiroshi TAKEUCHI
高崎健康福祉大学健康福祉学部医療福祉情報学科教授. 1971 早稲田大学大学院理工学研究科修士課程修了, 1976 理学博士．医療機器・情報システムの研究・開発に従事．電子情報通信学会、日本データベース学会、IEEE(Senior Member)、New York Academy of Sciences

児玉 直樹　Naoki KODAMA
高崎健康福祉大学健康福祉学部医療福祉情報学科講師. 2004 長岡技術科学大学大学院博士課程修了, 博士（工学）. 医用画像処理・情報システムの研究・開発に従事．日本放射線技術学会、電子情報通信学会、電気学会、ME 学会

橋口 猛志　Takeshi HASHIGUCHI
東京大学大学院医学系研究科健康医科学創造講座助手. 1994 東京大学医学部保健学科卒業, 医療情報システムの研究・開発に従事．日本医療情報学会

林 同文　Doubun HAYASHI
東京大学大学院医学系研究科健康医科学創造講座助教授. 1992 金沢大学医学部卒業（MD） 2004 医学博士（東京大学）. 循環器内科学・臨床スポーツ医学・医療情報システムの研究開発に従事．日本循環器学会、日本内科学会、日本体育協会、日本オリンピック委員会