音声言語情報処理　4－14
(1994. 12. 16)

対話情報に基づいた韻律生成

宮原 進　　山下洋一　　溝口理一郎

大阪大学産業科学研究所

〒567 大阪府茨城市美穂ヶ丘8番1号

あらまし

一般に対話では、発話文が同じでも対話コンテキストによって韻律的特徴が異なる。そこで、高品質の対話音声を合成するためには、対話から得られる情報を利用して韻律を決定する必要がある。本報告では、対話情報を用いて基本周波数を修正する対話規則の生成と評価結果について述べる。対話コンテキストのない孤立発声文から得られる韻律規則を別の対話データに適用する。この時の誤差を対話情報の欠如によるものと考え、誤差データに対話情報を与えたものを例題として対話規則を生成する。得られた対話規則は、特に台本に基づいて行なった対話のデータに対して、有効であることが確認された。

和文キーワード　対話音声, 対話コンテキスト, 対話情報, 対話規則, 基本周波数

Generation of prosody based on dialog information

Susumu Miyahara　　Yoichi Yamashita　　Riichiro Mizoguchi

Institute of Scientific and Industrial Research, Osaka University

8-1, Mihogaoka, Ibaraki city, Osaka prefecture, 567 Japan

Abstract

In order to synthesize natural dialog speech, it's necessary to incorporate dialog information with the prosody generation. This report describes a method of generating dialog processing rules which adjust the fundamental frequency using dialog information. Rules generated from utterances without dialog context predicts the fundamental frequency for another set of utterances with dialog context. We assume that prediction errors are due to lack of dialog information and generated the dialog processing rules from cases of the prediction error with dialog information. The evaluation test verifies that the dialog processing rules are effective to the utterances in dialogs, especially those in dialogs based on a script.

英文 key words　Dialog speech, Dialog context, Dialog information, Dialog processing rules, Fundamental frequency

1 はじめに

人間とコンピューターとの間での自然な対話を実現するために、我々は問題解決器に汎用な音声入出力インターフェースの構築を目指している。この枠組においては、問題解決器は表層表現や韻律等を考慮せずに発話の意味内容を概念表現で記述し、対話管理部が対話コンテキストに応じてこれを修正するとともに必要な対話情報を抽出する [1,2]。そしてそれに基づいて音声合成システムが合成音を出力する。本稿では韻律決定のための、対話情報を利用した規則生成の手法とその評価結果について述べる。

2 対話コンテキストに基づく音声合成

2.1 対話音声合成

対話中での発話は、孤立発話される文の発話とは異なり、同じ意味内容の発話を行なう場合でも、それまでの対話履歴,発話する状況等によって表層文表現や韻律等が大きく変化する。例えば、観光案内システムが利用者に対して『京都で特に見たいものがあるかどうかを尋ねる』発話は、概念表現 [2] を用いて

見る ([$object(何),$place(京都),$wish,
$interrogative]).

と記述される。これは通常

『京都で何が見たいですか』

のような表現に変換されるが、この発話の前までの対話で京都についての話が行なわれていたとすると

『何が見たいですか』

というように『京都で』が省略された方が自然な発話となる場合もある。また逆に、いくつかの地名がそれまでに対話中に現れ、特に京都に注目する時には

『京都で何が見たいですか』

というように『京都』が強調されたりする。このように同じ意味内容を対話コンテキストに応じて適当な音声言語表現に変換するには、対話から得られる情報を利用して表層文表現や韻律を決定するメカニズムが不可欠である。これまでに対話コンテキストによる表層文表現の多様性について検討した [3] が、本報告では対話音声合成において対話コンテキストに基づいて韻律を決定する手法について述べる。

2.2 対話規則生成の手法

対話コンテキストによって生じる韻律の変化をとらえるには、対話コンテキストのある発話とコンテキストのない発話とを比較してみるのがよいと思われ、この比較によって対話規則を生成し、対話音声の韻律決定を行なう。この対話規則は次のような手順により生成する [4]。

step1：対話コンテキストのないデータから合成規則を生成する。この規則を基本規則と呼ぶ。

step2：基本規則を対話コンテキストのあるデータに適用する。

step3：step2 の結果生じた誤差データに対話情報を与えたものを例題とし、これから再度規則を生成し、対話規則とする。

この対話規則を用いることにより、対話コンテキストに応じた韻律の決定を行なう。

3 音声データと F_0 モデル

今回用いた音声データは、対話コンテキストのない孤立発話としては ATR503 文を、対話コンテキストのある発話としては模擬対話 osa0010(以下対話データ I), osa1002 及び osa1003(以下対話データ II) の 2 種類のデータを用いた。ここで、対話データ I は二名の対話協力者に役割･目的を与えて行なう、いわゆる模擬対話である。対話データ II は台本に基づいて行なった対話であり、対話者は文表現された台詞にしたがって対話を行なった。この対話では、『必ずしも台詞通りに発話する必要はなく、若干の表現の変更を認める』条件で行ない、言い直し,言い淀みなども許してはいるものの、対話データ I に比べるとかなり読み上げに近くなっている。対話データとして用いた話者は一名のみであり、ATR503 の発話者と同一話者である。

対象とした韻律パラメータは基本周波数である。基本周波数のモデルとしては、阿部らが提唱する 2 階層制御方式 [5] を採用し、この方式のうちグローバルモデル、すなわちアクセント成分中の最大基本周波数のみを対象とした。また、この値は視察により決定した。

用いた例題数, すなわちアクセント成分の数は対話コンテキストのないデータについては 3288 個、対話コンテキストのあるデータについては対話データ I,II、それぞれ 250 個,220 個である。

4 基本規則の生成

4.1 基本規則の生成と評価

まず、対話コンテキストのないデータから基本規則を作成した。ここでの手法は阿部らのものと同じである [5]。すなわち、グローバルモデル (基本周波数のローカルピーク) に影響を及ぼすと思われる質的説明要因 (これを属性と呼ぶことにする) から、数量化I類によってグローバルモデルを決定する。従って、規則は数量化I類のモデルパラメータとして学習される。用いた質的説明要因も阿部らのものと同じである。

この規則をコンテキストのないデータに適用すると、表 1(1) のように平均誤差は 10.0Hz であった。この評価は、分割数 10 の CrossValidation によって行なった。この基本規則の重相関係数は 0.784 であり、阿部らの結果で報告されている 0.843 に比べるとやや低くなっている。これは、阿部らの話者がアナウンサーであるのに対し*、今回用いた話者は一般話者であり、発声のばらつきがやや大きいためと思われる。

次にこの基本規則をコンテキストのあるデータに適用すると、その平均誤差は、対話データIについては表 1(2) のように 20.1Hz, 対話データ II については表 1(3) のように 17.4Hz となった。この (1) と (2) の差, (1) と (3) の差が対話コンテキストの影響によるものといえる。なお参考までに、対話コンテキストのあるデータから作成した規則を対話コンテキストのあるデータに適用した結果を表 1(4),(5) に示す。この評価はクローズな評価である。

表 1: 基本規則を適用した時の結果

	学習データ	評価データ	平均誤差 [Hz]
(1)	ATR503 文	ATR503 文	10.0
(2)	ATR503 文	対話データ I	20.1
(3)	ATR503 文	対話データ II	17.4
(4)	対話データ I	対話データ I	18.7
(5)	対話データ II	対話データ II	14.7

*私信による

4.2 誤差データの定性的解析

対話情報を利用した韻律の決定を行なうためにはまず、どのような情報に着目すればよいかを知る必要がある。そこで、基本規則を対話コンテキストのあるデータに適用した際の誤差データを定性的に解析した。その結果、以下のような特徴が見受けられた。

基本周波数が増加する例

- 話題の転換時
  『あと･･･』, など.
- 新しい単語が出てきた時
- 挨拶などの時
  『はい、･･･です』, など.
- 驚きなどの感情の入った発話の時
- 前発話内容に情報を追加している時
  『それと･･･です』, など.

基本周波数が減少する例

- 前発話と同じ内容の時
- 相槌などの時
  『はい』, など.
- 落胆などの感情の入った発話の時
- 挨拶などの発話の時

5 対話規則の生成

先に述べたように、基本規則を対話コンテキストのあるデータに適用した結果の誤差データに対話情報を与えて例題とし、対話規則を作成する。

5.1 属性

4.2 での解析結果から、次のような対話に関する情報を例題の属性として与える。なお、個々の例題における属性値は人手によって決定した。

AT1: アクセント成分に新出語が含まれているかどうか (新固有名詞, 新出語, no)

AT2: 不要語の直後かどうか (yes,no)

AT3: 不要語の直前かどうか (yes,no)

AT4: 感情を含んでいるかどうか (戸惑い, 驚き, 落胆,no)

AT5: 前発話との関係 (展開, 詳細化, 継続, 回答, 相槌, 繰り返し, 確認)

AT6: 発話内での役割 (回答の中心, 追加, 変更, その他)

AT7: 各アクセント句における自立語の品詞 (10 カテゴリ)

これらの属性には、発話単位で決められるものとアクセント句単位で決められるものがある。発話単位で決められる属性では、その発話に含まれるすべてのアクセント句に同じ属性値が与えられる。発話単位で決定したものは AT4,AT5 であり、これら以外はすべてアクセント句単位で決定した。このことから、例えば 2 つの例題が同じ『展開』という属性値を持つと同時に、それぞれ『回答の中心』,『追加』という異なる属性値を持つこともある。これら 7 属性を用いて誤差データを説明する対話規則を生成する値を決定するための手法としては、数量化 I 類及び SBRtree を用いた。

5.2 SBRtree

属性から値を決定する手法としては数量化 I 類の他に SBRtree と呼ばれる手法も用いた。SBRtree を用いた規則の生成は、次のようなアルゴリズムによって行なう [6,7]。

step1 : まず、学習例題をルートノードへ割り当てる。

step2 : 次にある一つの属性に着目してノードを分割する。そして、リーフノードに達するまでこれを繰り返して決定木を作成する。

step3 : 決定木を作成した後、ルートノードからリーフノードまでの最適パスを選択する。本報告では最適なパスとしてリーフノードまでのパスが最短であるものを用いた。

step4 : step3 で選択された最適パス上のノードにある条件のセットを規則として抽出し、リーフノードに含まれる例題の基本周波数の平均をその規則に対する基本周波数成分として割り当てる。

step5 : リーフノードに含まれる例題を学習例題セットから削除する。

step6 : step5 の結果残った例題を、再びルートノードに割り当て、step2 へ。残った例題がなければ終了。

5.3 数量化 I 類による対話規則生成

5.3.1 対話規則の評価

数量化 I 類によって生成した対話規則を対話データ I, 対話データ II に適用した。その結果をそれぞれ表 2(a),(b) に示す。なお、オープンな評価は 10 分割の CrossValidation によって行なった。表からわかるように、全誤差データに対しての誤差の減少値は 1Hz 前後であった。しかし、すべての発話、あるいはすべてのアクセント成分 (文節) が必ずしも対話コンテキストの影響を受けるわけではない。対話コンテキストの影響を大きく受けているところでは、基本規則を適用した時の誤差が大きいと考えられる。このことから、誤差が 30Hz 以上あったデータのみに対して対話規則を適用してみた。その結果が表の E30 である。なお、30Hz 以上の誤差があった例題数は対話データ I,II、それぞれ 57 個,33 個である。このような誤差の大きいデータに対しては、対話データ I については誤差が約 4Hz しか減少しなかったが、対話データ II については 10Hz 以上減少した。

表 2: 数量化 I 類による対話規則の評価

(a) 対話データ I

評価対象	対話規則適用前 [Hz]	規則適用後 [Hz]	
		クローズ	オープン
(ALL)	20.1	17.2	19.6
(E30)	41.4	33.2	37.1

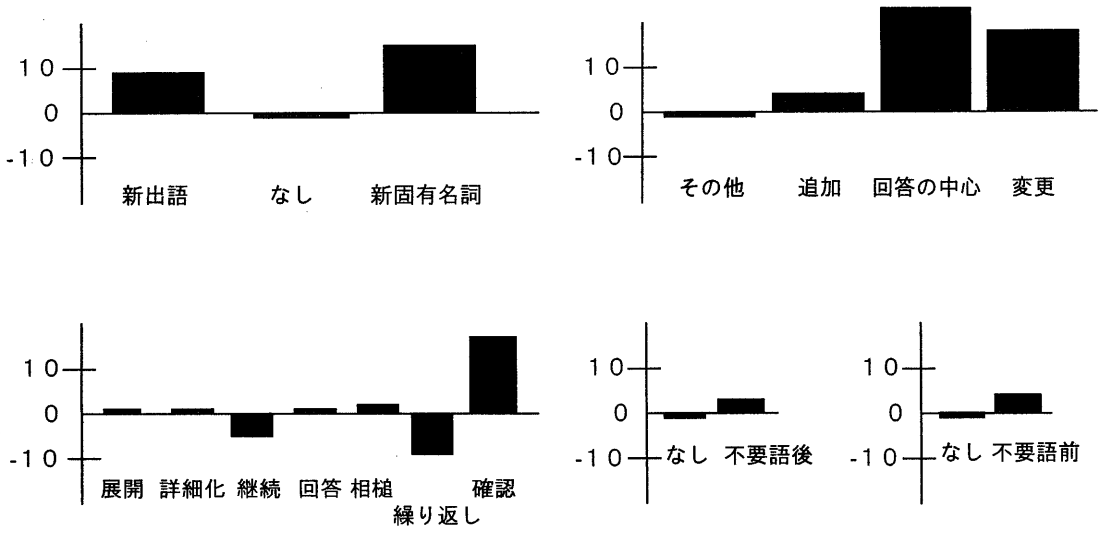
(b) 対話データ II

評価対象	対話規則適用前 [Hz]	規則適用後 [Hz]	
		クローズ	オープン
(ALL)	17.4	13.6	16.2
(E30)	38.9	25.1	27.2

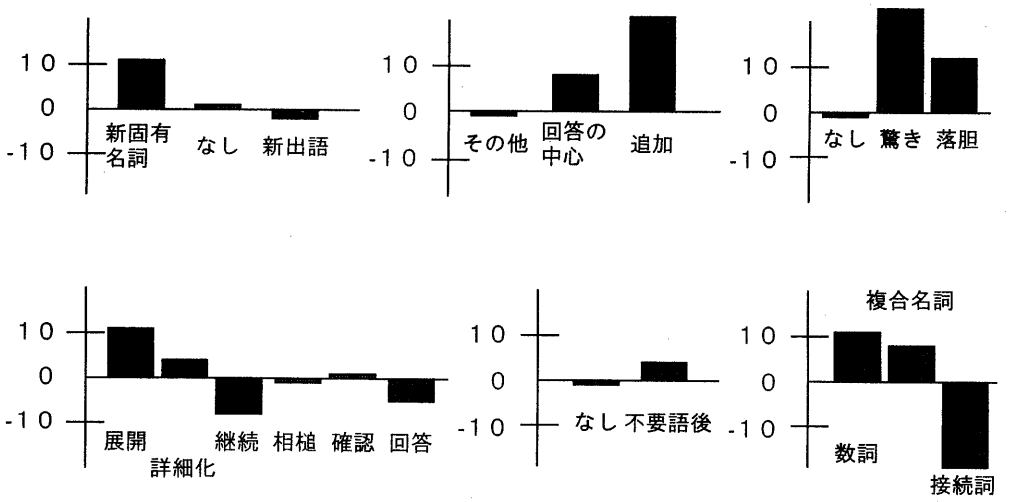
(ALL) : 全誤差データ
(E30) : 誤差が 30Hz 以上あったデータ

5.3.2 対話規則の特徴

対話規則として得られた数量化 I 類のモデルパラメータから、図 1 のような特徴が見受けられた。図 1(a) からわかるように、対話データ I については、新出語では基本周波数が増加している。また、不要語の直前、あるいは直後で増加している。前発話との関係においては、展開, 詳細化, 確認のところでは増加し、逆に前発話と同じ内容の繰り返しのところでは減少している。さらに、発話内で回答の中心, 追加, 変更の位置にあるところでは増加しているのがわかる。対話データ II の方についても同様な特徴が見受けられた。この中でも特に、数詞や複合名詞では増加し、逆に接続詞で減少しているという特徴が見られた。しかし、対話データ I の時とは異なり、固有名詞以外の新出語では減少していた。これは、対



(a) 対話データ I



(b) 対話データ II

図 1: 数量化 I 類による対話規則の特徴

話データ II が 2 種類の模擬対話における同一話者部分を合成して作成されたものであり、この 2 種類の模擬対話のうち、平均基本周波数の低い方でこういった単語が多く発せられているためであると思われる。その他の特徴については、先の解析結果とほぼ一致していた。

5.4 SBRtree による対話規則生成

5.4.1 対話規則の評価

SBRtree によって生成した対話規則を、対話データ I, 対話データ II に適用した結果をそれぞれ表 3(a),(b) に示す。なお、オープンな評価は 10 分割の Cross-Validation によって行なった。

SBRtree によって生成された対話規則の評価では、数量化 I 類による規則とほぼ同様の結果が得られた。このうち、対話データ I を用いた結果では対話情報の効果が期待していたほどは得られなかった。この理由として、対話データ I ではシステム側が愛想のない応答を返すように意図的に設定してあったため、スムーズさに若干欠けるものになってしまっていたことが考えられる。また、実際の対話ではいろいろな心的要因が絡み合っての発話が多く、今回用いた対話情報だけでは説明しきれなかったと考えられる。一方対話データ II を用いた結果では、全誤差データに対する誤差の減少値は小さかったものの、基本規則による誤差が 30Hz 以上あったものについては減少値が大きく、対話データ I の結果と比べると、対話規則による効果がかなり大きい。これは、対話データ II がやや読み上げに近い対話であるため、複雑な心的要因等の影響が少ないためと思われる。

5.4.2 対話規則の特徴

SBRtree によって生成した対話規則の例を表 4 に示す。なお、括弧内の数字はそれぞれの規則が学習時にカバーした例題数である。対話データ I では、学習例題が特定の規則に偏ってカバーされ、うまく学習されているとは言い難いものの、『戸惑った時に基本周波数が増加する (rule2)』など、納得できるものも見られた。対話データ II については、対話データ I に比べると学習例題が平均的にカバーされており、『新出語で基本周波数が増加する (rule4)』など、かなり納得のいくものも得られた。

5.5 適用結果の具体例

以上の適用結果の具体例を次に挙げる。

表 3: SBRtree による対話規則の評価

(a) 対話データ I

評価対象	対話規則適用前 [Hz]	規則適用後 [Hz]	
		クローズ	オープン
(ALL)	20.1	17.3	19.6
(E30)	41.4	35.2	38.5

(b) 対話データ II

評価対象	対話規則適用前 [Hz]	規則適用後 [Hz]	
		クローズ	オープン
(ALL)	17.4	13.4	15.7
(E30)	38.9	26.2	30.4

(ALL) : 全誤差データ
(E30) : 30Hz 以上の誤差があったデータ

まず、回答の中心であるところで増加する例である。

システム：『バスで行く行き方とジェーアールで行く行き方とがありますが。』
ユーザ　：『ジェーアールの場合どうなるんでしょうか。』
システム：『ジェーアールですと…』

このユーザ側の発話における下線部の基本周波数は、対話コンテキストのある中で発話された時には 222Hz であったが、基本規則を適用した結果 190Hz と決定された。これに『回答の中心』という対話情報を与えたところ、218.9Hz まで改善された。

次に、前発話と同じ話題の継続である発話において減少する例である。

システム：『…駅前のデパートの 5 階にも温泉があります。』
ユーザ　：『デパートに温泉があるんですか。』
システム：『はい。他には上諏訪駅の近くに湖畔公園や、…』

このシステム側の発話における下線部では、対話コンテキストのある中で発話された時には 177Hz であったが、基本規則を適用した結果 190.3Hz と決定された。これに『継続』という話題情報を与えたところ、この値は 176.2Hz まで改善された。

これらの例は、対話情報を与えることにより基本周波数の値が改善されたものであるが、中には次のようにかえって誤差が増加する例も見られた。

表 4: SBRtree による対話規則の例

(a) 対話データ I

rule1: 【新出語 & ～不要語後 & 不要後前】	→	−6Hz	(2)
rule2: 【～不要語後 & 戸惑い & (名詞 | 動詞)】	→	41.2Hz	(2)
rule3: 【～不要語後 & 変更】	→	21.6Hz	(2)
rule4: 【不要語後 & 不要語前 & 副詞】	→	54.7Hz	(2)
rule5: 【～新出語 & ～不要語後 & ～感情】	→	−8.6Hz	(148)
rule6: 【～新出語 & ～不要語前 & 展開 & ～役割 & (動詞 | 名詞)】	→	1.9Hz	(18)
rule7: 【不要語前】	→	6.4Hz	(20)
rule8: 【不要語後 & 展開 & (動詞 | 名詞)】	→	13.5Hz	(9)
rule9: 【新出語 & (繰り返し | 展開)】	→	−10.6Hz	(9)
rule10: 【～新出語】	→	4.4Hz	(27)
rule11: 【】	→	15.3Hz	(12)

(b) 対話データ II

rule1: 【(～新出語 | 新出語) & ～不要語後 & (継続 | 回答) & ～役割 & (名詞 | 数詞 | 複合名詞 | 形式動詞 | 形容動詞)】	→	−12.7Hz	(52)
rule2: 【～不要語前 & (詳細化 | 回答) & (～役割 | 回答の中心) & (副詞 | 名詞 | 複合名詞)】	→	−0.4Hz	(38)
rule3: 【～不要語前 & (～感情 | 落胆) & (詳細化 | 継続 | 回答) & (動詞 | 連体詞 | 接続詞 | 形容詞 | 感動詞)】	→	−20.1Hz	(25)
rule4: 【(新固有名詞 | 新出語) & 展開】	→	10.3Hz	(19)
rule5: 【～不要語前 & ～感情 & ～役割 & (形式名詞 | 数詞 | 名詞 | 動詞)】	→	3.5Hz	(37)
rule6: 【～新出語 & ～不要語後 & ～感情】	→	−16.8Hz	(31)
rule7: 【～不要語後 & (詳細化 | 継続)】	→	20.1Hz	(7)
rule8: 【～感情】	→	6.6Hz	(6)
rule9: 【】	→	23.4Hz	(5)

ユーザ　：『電車の本数は結構ありますか。何分に一本くらいあるんでしょうか。』
システム：『ジェーアールは10分に1本ぐらいです。』
ユーザ　：『大田電鉄の方は』
システム：『8分に一本ぐらいです。』

このユーザ側の発話における下線部では、対話コンテキストのある中で発話された時には168Hzであったものは、基本規則を適用した結果128.5Hzと決定され、対話情報を与えたところ、103.5Hzとなり改善はされなかった。

このように、対話情報を与えても基本周波数の値が改善されていないものが他にもあり、今後新たな対話情報を考慮していく必要があるように思われる。

6　おわりに

対話コンテキストのない発話から生成した規則を、対話コンテキストのある発話に適用することによって生じる誤差を利用し、対話情報を利用した韻律規則の生成を行なった。今回の実験では、対話コンテキストのある発話データとして性質の異なる2種類のものを用いた。このうち、スムーズさに若干欠けている発話データでは対話情報による効果が期待していたほどは得られなかったが、これは基本周波数のばらつきが大きいために、今回用いた対話情報だけでは説明できなかったためであるように思われる。それに対し、やや読み上げに近くなっている発話データでは対話情報の効果がよく得られたと言える。なお、今回の実験では評価データ数が十分であるとは言えないため、新たな対話特徴の抽出を行いながら実験データを増やしていく予定である。

謝辞

本研究の一部では、文部省科研費(重点領域研究『音声対話』, No.05241105)の援助を受けた。なお、本研究の際に御協力頂いた大阪大学大学院工学研究科の作田 瑞氏に感謝致します。

参考文献

[1] Yamashita, Y., et al, "MASCOTS II: A Dialog Manager in General Interface for Speech Input and Output", IEICE Trans., vol.E76-D, no.1, pp.74-83(1993)

[2] 山下洋一, 他:" 汎用音声出力インタフェースにおける概念表現からの音声合成", 信学論,J76-D-II, No.3 pp.415-426(1993)

[3] 田島慶一, 他:" 対話コンテキストを利用した概念表現からの対話文生成", 人工知能学会研究会,SIG-SLUD-9302-9,pp.65-72(1993)

[4] 宮原 進, 他:" 対話音声における韻律の特徴抽出と規則化", 音講論,2-5-5,pp.275-276(1994)

[5] 阿部匡伸, 他:" 音節区分化モデルに基づく基本周波数の2階層制御方式", 音響学会誌,49,10,pp.682-690(1993)

[6] Yamashita, Y., et al, "Tree-Based Approaches to Automatic Generation of Speech Synthesis Rules for Prosodic Parameters", IEICE Trans., vol.E76-A, no.11, pp.1934-1941(1993)

[7] Indurkhya, N., et al, "Iterative Rule Induction Methods", Applied Intelligence, vol.1, pp.43-54(1991)