

4Q-7

事例間の局所的な関係と分類の大域的な概観からの
直観的クラスタリングの誘導

斉藤 康彦

東条 敏

情報処理振興事業協会

1 概要

事例からの概念記述の誘導は、機械学習の領域で研究されてきた。多くの概念帰納学習の研究では、事例や概念の記述に述語論理が用いられている [2]。概念クラスタリングのプログラムである CLUSTER/S では、構造的記述を取り扱えるように、述語論理が拡張されている [4]。ID3 という帰納アルゴリズムでは、概念は、決定木として記述される [3]。ひとつの事例は、節が属性に対応し、枝が属性値に対応するような決定木における、ひとつの葉に割り当てられる。属性に基づく表現言語は、事例を特徴付ける記述が首尾一貫していること、および、カテゴリを特徴付ける概念が事例の一般化であることを前提にしている。

一方、文芸評論家が、ある作家の全作品を「重い」作品と「軽い」作品に分類するというような状況では、事例や概念の記述に、述語論理や決定木を用いることは困難である。彼は、ある作品が他の作品と似ているということ直観的に判断できるし、「重い」「軽い」という言葉の意味を直観的に理解できる。それにもかかわらず、ある作品が「重い」か「軽い」かを、機械的に決定することは難しい。なぜならば、各作品を特徴付ける属性や、「重い」「軽い」というカテゴリの一般的な記述が、明白ではないからである。たまたま見出しされた属性も、すべての作品を首尾一貫して説明できないかもしれないし、目標とするカテゴリに関係しているとはかぎらない。

このような状況は、直観的分類の枠組を必要とする。事例を特徴付ける記述が、互いに矛盾していたり、カテゴリを特徴付ける概念が、その概念に属する事例と矛盾していたりすることを、許容する枠組を考えなければならない。本論文で提示する直観的分類の基本メカニズムでは、分類の大域的な概観が、事例間の局所的な関係の統一性の欠如を埋め合わせる。

しばしば普通の辞書の代わりに用いられるシソーラスからの類推によって、ある事例は、その事例と類義的または反義的な事例の集合として、直観的に記述されると仮定する。局所的な類義性/反義性のような関係を取り扱うことは、理にかなっている。これは、ふたつの事例が類義的であるか反義的であるかを、直観的に判断することができるからである。しかし、他の多くの種類の関係を取り扱うことは、理にかなっていない。これは、それらの間の区別が難しくなるからである。

完全なシソーラスは、完全な語の定義から導かれ、完全な語の分類は、完全なシソーラスから導かれる。逆に言えば、事例間の局所的な関係から、有意義なクラスタリングが得られるとはかぎらない。このような場合には、分類の意図を与えることによって、有意義なクラスタリングが得られる。分類の意図は、典型的な事例を含み、カテゴリを暗黙的に定義するクラスターの集合として、直観的に与えられると仮定する。このようなクラスターの集合は、望ましいクラスタリングの大域的な概観を間接的に示している。ここで、あるカテゴリは、それに属するための必要十分条件の集合として定義されない [1]。カテゴリは、典型的な事例と非典型的な事例を含んでいる。

本論文で提示する直観的分類の枠組では、事例のクラスタリングは得られるが、分類規則は生成されない。学習システムが分類規則を生成することは、理にかなっている。しかし、直観的な情報を取り扱う問題では、分類規則を得ることは困難である。すべての事例が適切に分類されているならば、事例のクラスタリングは、問題解決のために、必須であり、かつ、有用である。

2 親和性に基づくクラスタリング

親和性に基づくクラスタリングは、語を語群に分類するための手法である。領域は、分類するすべての語の集合である。各語について、類義語と反義語が定義される。語 w の類義語の集合を $w.Syn$ と表記する。語 w の反義語の集合を $w.Ant$ と表記する。これらの集合は、領域の部分集合である。見本は、典型的な語からなる語群の集合である。見本の中の各語群は、領域の部分集合である。したがって、直観的分類の問題は、以下のように記述される。

ある領域と見本が与えられた時に、領域の中のすべての語の、排反なクラスタリングを求めよ。ただし、このクラスタリングは、見本に準拠していなければならないものとする。

ここで注意すべき点は、生成されるクラスターの判別記述（換言すれば、分類規則）が、分類のために用いられないことである。

与えられた領域の中の各語は、与えられた見本の中のある語群に割り当てられる。この語群が、その語の最適クラスターである。ある語の最適クラスターは、親和性に基づいて決定する。語 w と語群 C の間の親和性の定義を、次頁上段に示す。

ある語とその語の最適クラスター候補のひとつの間の親和性は、その語とそれ以外の語群の間の親和性に等しいか、または、より大きい。しかし、ある語とその語の最適クラスターの間の親和性は、その語とそれ以外の語群の間の親和性より大きい。そこで、ある語の最適クラスター候補は、必ず決定するが、ある語の最適クラスターは、必ずしも決定しないことに、注意しなければならない。

3 クラスタリングの再編

クラスタリングの再編は、領域の中の語のすべてについての、最適クラスターを決定するための手続きである。その手続きは、分配と変形からなる。分配とは、最適クラスターの条件を緩和しながら、領域の中のすべての語を最適クラスターに分配することである。分配のアルゴリズムを、次頁中段に示す。ある語の最適クラスターが、それでもなお、決定しないならば、その語はどのクラスターにも属さないまま残される。変形とは、残された語の少なくともひとつを除去するために見本を変形することである。変形のアルゴリズムを、次頁下段に示す。分配と変形は、残された語がなくなるまで反復される。

最終的なクラスタリングは、与えられた見本に準拠していなければならない。換言すれば、見本を変形しすぎてはならない。このために、以下のような戦略で再編を行なう。

1. 残された語のひとつについての最適クラスターが、変形によって必ず決定するようにする。
2. ある語の最適クラスターが、変形前に既に決定しているならば、変形後も決定するようにする。

したがって、残された語の数は、再編によって減少する。

謝辞

本研究は、次世代産業基盤技術研究開発「新ソフトウェア構造化モデルの研究開発」の一環として情報処理振興事業協会が新エネルギー・産業技術総合開発機構から委託をうけて実施したものである。

参考文献

- [1] Holland, J.H., Holyoak, K.J., Nisbett, R.E. and Thagard, P.R.: *Induction: Processes of Inference, Learning, and Discovery*, MIT Press (1986).
- [2] Michalski, R.S.: A Theory and Methodology of Inductive Learning, in Michalski, R.S., Carbonell, J.G. and Mitchell, T.M. (Eds.), *Machine Learning: An Artificial Intelligence Approach*, Tioga (1983).
- [3] Quinlan, J.R.: Learning Efficient Classification Procedures and their Application to Chess End Games, in Michalski, R.S., Carbonell, J.G. and Mitchell, T.M. (Eds.), *Machine Learning: An Artificial Intelligence Approach*, Tioga (1983).
- [4] Stepp III, R.E. and Michalski, R.S.: Conceptual Clustering: Inventing Goal-Oriented Classifications of Structured Objects, in Michalski, R.S., Carbonell, J.G. and Mitchell, T.M. (Eds.), *Machine Learning: An Artificial Intelligence Approach* (Vol.2), Morgan Kaufmann (1986).

語 w と語群 C の間の親和性の定義

$$\text{affinity}(w, C) = \text{assoc1}(w, C) + \text{assoc2}(w, C) + \text{assoc3}(w, C) + \text{assoc4}(w, C)$$

where

$$\begin{aligned} \text{assoc1}(w, C) &= \begin{cases} 1 & \text{when } \exists x : x \in w.Syn \wedge x \in \bigcup_{a \in C} (a.Syn) \\ 0 & \text{when } \sim \exists x : x \in w.Syn \wedge x \in \bigcup_{a \in C} (a.Syn) \end{cases} \\ \text{assoc2}(w, C) &= \begin{cases} -1 & \text{when } \exists x : x \in w.Syn \wedge x \in \bigcup_{a \in C} (a.Ant) \\ 0 & \text{when } \sim \exists x : x \in w.Syn \wedge x \in \bigcup_{a \in C} (a.Ant) \end{cases} \\ \text{assoc3}(w, C) &= \begin{cases} -1 & \text{when } \exists x : x \in w.Ant \wedge x \in \bigcup_{a \in C} (a.Syn) \\ 0 & \text{when } \sim \exists x : x \in w.Ant \wedge x \in \bigcup_{a \in C} (a.Syn) \end{cases} \\ \text{assoc4}(w, C) &= \begin{cases} 1 & \text{when } \exists x : x \in w.Ant \wedge x \in \bigcup_{a \in C} (a.Ant) \\ 0 & \text{when } \sim \exists x : x \in w.Ant \wedge x \in \bigcup_{a \in C} (a.Ant) \end{cases} \end{aligned}$$

分配のアルゴリズム

現在の語を w とする.

n 個の語群を含む見本を $SMPL$ とする.

各 $C_i \in SMPL$ について,

$\text{sum}(C_i) := 0$;

各 $s \in w.Syn$ について,

IF $\forall x \in SMPL : \text{affinity}(s, C_i) \geq \text{affinity}(s, x)$, THEN $\text{sum}(C_i) := \text{sum}(C_i) + 1$;

各 $a \in w.Ant$ について,

IF $\forall x \in SMPL : \text{affinity}(a, C_i) \leq \text{affinity}(a, x)$, THEN $\text{sum}(C_i) := \text{sum}(C_i) + 1$.

IF $\text{sum}(C_i) > (\sum_{k=1}^n \text{sum}(C_k))/2$, かつ, C_i は、 w の最適クラスター候補である, THEN

C_i は、 w の最適クラスターである.

ELSE

IF 現在の語の最適クラスター候補のひとつは、現在の語の変形前の最適クラスターの部分集合である, THEN
当該最適クラスター候補は、現在の語の最適クラスターである.

ELSE

IF 現在の語の変形前の最適クラスターは、現在の語の最適クラスター候補のひとつの部分集合である, THEN
当該最適クラスター候補は、現在の語の最適クラスターである.

ELSE

IF あるクラスターは、現在の語のみを要素とする語群である, THEN
当該クラスターは、現在の語の最適クラスターである.

ELSE

現在の語の最適クラスターは、決定しない.

変形のアルゴリズム

現在の語を w とする.

w の最適クラスター候補を C とする.

IF C および w は以下を満足する:

$\text{affinity}(w, \{x|w.Syn \cap x.Ant = \phi, x \in C\}) = \text{affinity}(w, C) + 1$,
ただし、 C を $\{x|w.Syn \cap x.Ant = \phi, x \in C\}$ で置換することになる; または
 $\text{affinity}(w, \{x|w.Ant \cap x.Syn = \phi, x \in C\}) = \text{affinity}(w, C) + 1$,
ただし、 C を $\{x|w.Ant \cap x.Syn = \phi, x \in C\}$ で置換することになる.

C を $\{x|w.Syn \cap x.Ant = \phi, x \in C\}$ で置換する時には,

$\forall y : (y \in G(C)) \rightarrow (\text{affinity}(y, \{x|w.Syn \cap x.Ant = \phi, x \in C\}) \geq \text{affinity}(y, C))$.

C を $\{x|w.Ant \cap x.Syn = \phi, x \in C\}$ で置換する時には,

$\forall y : (y \in G(C)) \rightarrow (\text{affinity}(y, \{x|w.Ant \cap x.Syn = \phi, x \in C\}) \geq \text{affinity}(y, C))$.

THEN

C を $\{x|w.Syn \cap x.Ant = \phi, x \in C\}$ a または $\{x|w.Ant \cap x.Syn = \phi, x \in C\}$ で置換する.

ELSE

w の最適クラスター候補を C_1, C_2 とする.

IF C_1, C_2 および w は以下を満足する:

$\text{affinity}(w, C_1 \cup C_2) = \text{affinity}(w, C_1) + 1 = \text{affinity}(w, C_2) + 1$,

特に、最適クラスター候補の数が2である時には,

$\text{affinity}(w, C_1 \cup C_2) = \text{affinity}(w, C_1) + 1 = \text{affinity}(w, C_2) + 1$ または

$\text{affinity}(w, C_1 \cup C_2) = \text{affinity}(w, C_1) = \text{affinity}(w, C_2)$.

$\forall x : (x \in G(C_1)) \rightarrow (\text{affinity}(x, C_1 \cup C_2) \geq \text{affinity}(x, C_1))$ かつ

$\forall x : (x \in G(C_2)) \rightarrow (\text{affinity}(x, C_1 \cup C_2) \geq \text{affinity}(x, C_2))$.

THEN

$C_1 \cup C_2$ を見本に追加し、 C_1 と C_2 を見本から削除する.

ELSE

現在の語のみを要素とする語群を見本に追加する.