平成13年度卒業研究		
入学年度　平成9年度	学籍番号　09117913	氏名　　小川 智也
論文題目　事例間類似度計算の特徴選択による効率化		犬塚研究室

1　はじめに

データベースに含まれるオブジェクト間の類似性(距離)を数値化することによって、データベース中から様々なパターンや一定の規則を見つけることができる。

本論文ではデータベース中の事例や属性などの類似度を測るICD (Iterated Contextual Distance)アルゴリズムに注目し、その計算量が大きいことを改善する手法として、**特徴選択**の手法を提案する。特徴選択(属性選択)は、分類問題などに関係のない属性や冗長なものをを取り除く問題であり、データクリーニングの重要な一部である。

2　ICDアルゴリズム

ICDアルゴリズム[1]は事例間、属性間、部分関係(同じ属性を持つ事例の集合)間の距離を互いに関連付けて尤もらしい事例間の距離を求めている。概要は以下の通り。

1. 属性間の距離を乱数で初期化
2. 収束するまで3〜5を繰り返す
3. 属性間の距離から事例のベクトル表現を導く
4. ある属性を含む事例集合(部分関係)間の距離をそこに含まれる事例ベクトル間の距離として求める
5. 属性とそれを含む事例の集合(部分関係)を同一のものとし、属性間の距離を求める

ICDアルゴリズムは、ある属性を持つ事例集合が他にどのような属性を持っているかにより、異なる属性の事例集合間との類似度を求めている。事例集合が持つ属性の持ち方が類似していれば、その事例集合間は類似していると見なされる。この手法は事例数m、属性数nに対し、漸近的計算量が$O(mn^2)$となる。

3　提案手法：関連度による属性の選択

本研究では属性間の関連を提案し、それに基づいて関連の低い属性を不必要な属性として取り除き、計算量を下げる方法を提案する。属性A_iと他の属性との関連の度合(関連度)を表すベクトル$\mathbf{V}_{A_i}$を以下のように定義する。

$$\mathbf{V}_{A_i} = \mathbf{S}_{A_i} + \mathbf{E}_{A_i} \quad (i = 1, .., n)$$

ここで、$\mathbf{S}_{A_i}$はベクトルであり、そのj番目の要素はA_iと$A_j(j = 1, .., n)$が共に持っている事例の個数を表す。この値が0でない時、A_iとA_j間に1次的な関連が有ると呼ぶ。また、ベクトル$\mathbf{E}_{A_i}$のj番目の値はA_iと全ての$A_k(k = 1, .., n)$の1次的な関連度の値にA_kとA_jの1次的関連の値を掛けたもので、A_kから見た場合のA_iとA_jの類似性を表し、これを等価度を呼ぶ。本論文では$\mathbf{V}_{A_i}$がある閾値以上の要素を含む属性を選択してICDアルゴリズムを適用することを提案する。

4　実験結果

図1では、本学の教官データベース(事例数:429, 属性数:549)に対し、ICDアルゴリズムと提案手法による特徴選択を組み合わせた場合の属性数、計算時間、結果の一致度の推移を表す。実験結果の具体的な例を知能情報システム学科の教授を挙げて示す。表1のキーワードを含むデータベースから求めた距離に基づき図2を作成した。図1の一致度は、全ての教官に対して、その教官と最も類似する教官上位5人が集合として一致する割合を求めた。グラフの推移を見ると、閾値4において、不要な属性を取り除くことで計算時間を大幅に減らしながらも一致度の低下は少ない。それ以上の閾値では必要な属性も取り除かれ、一致度の低下を招いたと考えられる。

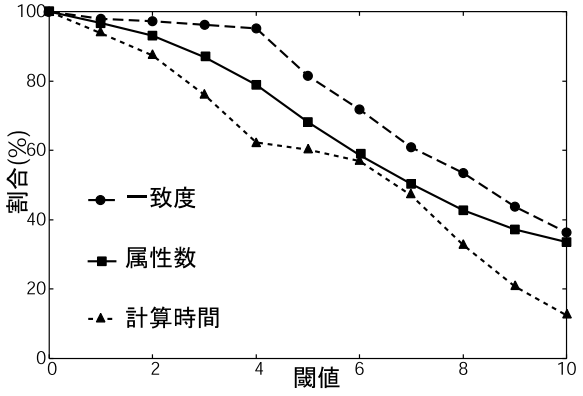

図1: 閾値による計算時間、属性数、一致度の推移

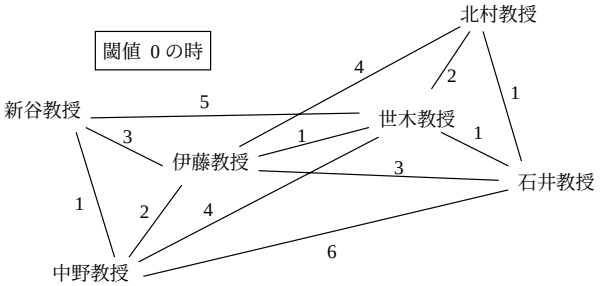

図2:ICDアルゴリズムで求めた事例間距離の例

表 1: 知能情報の教授の研究キーワード

石井教授	情報 工学 人工 知能 処理 ニューラル ネットワーク 生体 情報 処理
伊藤教授	人工 知能 プログラム 言語
北村教授	音声 情報 処理 画像 情報 処理 マルチメディア 情報 処理
新谷教授	知識 工学 人工 知能 意思 決定論 論理型 プログラミング ソフトウェア 工学
世木教授	計算機 科学 人工 知能
中野教授	人工 知能 神経 回路網 進化 計算

5　まとめ

本研究では 属性間の関連に基づいた属性選択手法の提案、実験を行なった。これにより、不要な特徴を取り除くことで計算量の減少に成功し、一致度の低下を抑えた。

参考文献

[1] Gautam Das and Heikki Mannila. “Context-Based Similarity Measures for Categorical Database.” PKDD 2000, pp. 201–210.