1S-8

クラスタ分析手法を用いた決定木学習手法の改善に関する研究

○天沼沙織† 榑松理樹† 羽倉淳† 藤田ハミド†
岩手県立大学大学院ソフトウェア情報学研究科†

１．はじめに

データマイニングの代表的な手法の一つに決定木学習手法[1]がある。これは、複数の属性をもつデータ集合から、一つの属性（目的属性）の値に対し、他の属性値の組み合わせから構築される分類規則を木の形で構築する手法である。結果の可読性や計算量の観点から多くの分野で活用されているが、推定精度が元になるデータに影響される性質がある。推定精度を高めるためには事前に高品質のデータ集合を用意する必要があるが、推定するデータが不明である点等から、その実施は困難である。この問題に対し、データを細分化する手法[2]、属性追加等の前処理を加える手法[3]等が取り組まれてきたが、問題の解決には至っていない。

一方、決定木の構築過程においては、属性間の関係を考慮していない。属性間の関係を考慮すれば、新たな観点でのデータ分割も可能となり、推定精度の向上が期待できる。本研究ではこの点に着目し、従来の決定木学習では利用していない属性間の関係も活用して決定木の構築を試みる。具体的には、従来の決定木学習手法に対し、非階層クラスタ分析[4]を用いた分割手法を追加することで、より細分化された決定木の構築を試みる。

２．クラスタ分析を用いた決定木の構築

２．１　学習方法

本研究で提案する決定木学習手法のアルゴリズムを図１に示す。本手法では、初めに、入力となるデータ集合のエントロピーと事前に設定する閾値とを比較する。エントロピーが閾値より大きい場合、従来の決定木学習手法に基づき入力データ集合を分割し、それぞれを子ノードとする。エントロピーとは情報の曖昧さを表す指標であることから、目的属性の値が偏っていない場合、子ノードを生成することになる。一方、エントロピーが閾値以下の場合、非階層クラスタ分析を用いてクラスタ集合を生成する。得られたクラスタ集合のクラス間分散[5]が事前に与えた閾値より大きい場合、得られたクラスタ集合を子ノードとする。クラス間分散は、クラス間のバラつきを表す指標であることから、一定以上距離があるデータに分割できる場合、子ノードを生成することとなる。一方、クラス間分散が閾値よりも小さい場合は入力データ集合をリーフノードとする。いずれかの方法で子ノードが得られた場合、それぞれを新たな入力データ集合とし、再帰的に処理を行う。

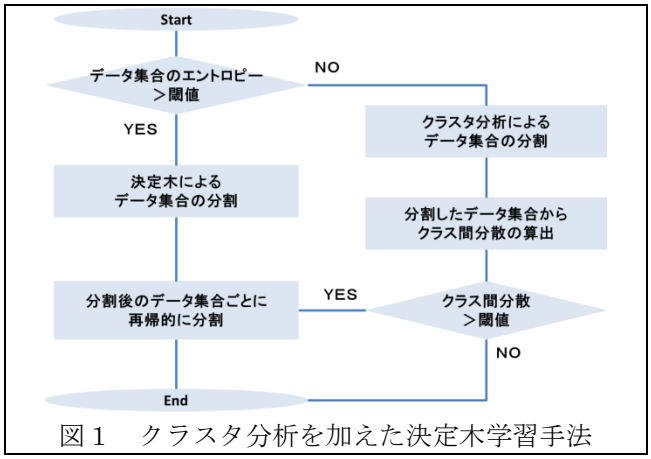

図１　クラスタ分析を加えた決定木学習手法

２．２　推定方法

本手法で生成された決定木は、従来手法では生成されないノードを含む。推定時にそれらのノードを用いる場合、従来とは異なる推定方法が必要となる。以下、その方法について説明する。

従来の決定木で生成されたノードには、一つの属性値に対する条件が記載されている。推定対象となるデータに対し、この条件を適用し、子ノードへと処理を進める。一方、クラスタ分析によって生成されたノードには、各クラスタの中央値が与えられている。そのため、この値とデータとの距離を求め、最近傍のクラスタになる子ノードへと処理を進める。これらの処理をリーフノードに達するまで繰り返す。そして、到達したリーフノードに含まれる目的属性の値が推定結果となる。

３．評価実験

３．１　実験概要

決定木の精度は与えるデータの影響を受ける点を考慮し、本提案手法の有用性および有用なデータ集合の特性を検証するために次に示す実験を行った。

(1) 実験データおよび手法：実験データは多様な特性をもつデータ集合を用意する必要性から、ランダムに生成した 20 個のデータ集合を用いる。各データ集合は４クラス各 50 個の計 200 個のデータから構成されており、各データは３つの属性を持つ。実験においては、各データ集合に対し、160 個（各クラス 40 個）のデータを決定木構築、残り 40 個（各クラス 10 個）を検証用

A Research on Decision Tree Learning Improvement Using Cluster Analysis
†Saori AMANUMA・Iwate Prefectural University, Software and Information Science

とし、10 回の試行からなる交差検査法を行う。

(2) 手法設定：実験においては、決定木学習には ID3[1]、クラスタ分析には k=2 と設定した k-means[6]を用いた。エントロピーの閾値を 0.5、0.75、1.0、1.25、1.5 とし、クラス間分散の閾値を 0.5、0.75、1.0 と設定した。

(3) 評価方法：提案手法の有用性の評価を行うために、同じデータに対する ID3 の処理結果と推定精度について比較を行う。ここで推定精度は、推定された結果が正解と一致した割合を意味する。また、両手法で生成された決定木の葉の数、深さの比較を行う。

３．２　実験結果・評価

初めに表１に両手法における推定精度の変動の一例を示す。ここで変動は、提案手法の推定精度÷ID3 の推定精度で求めている。データ集合#4 は常に向上したものであり、データ集合#15 は常に劣化したものである。全体としては、86 件のデータセットでは向上したが、214 件は劣化した。次に表２に両手法で生成された決定木について、葉の数と木の深さの増減を示す。表２においての上段が葉の数、下段が深さを示す。葉の数は増加したが、木の深さの変化が小さいことから、決定木手法では分割が不十分だった箇所の分割が進んだと考えられる。以上のことから、提案手法は、データ集合の特性によっては有用であると考えられる。

表１　推定精度の変動の一例

	データ集合#4			データ集合#15		
クラス間分散 / エントロピー	0.5	0.75	1.0	0.5	0.75	1.0
0.5	+1.0	+1.0	+1.0	0.9	0.9	0.9
0.75	+1.0	+1.0	+1.0	0.9	0.9	0.9
1.0	+1.1	+1.1	+1.1	0.9	0.9	0.9
1.25	+1.1	+1.1	+1.1	0.7	0.7	0.7
1.5	+1.1	+1.1	+1.1	0.6	0.6	0.6

表２　決定木の比較(空白は 0.0)

		データ集合#4			データ集合#15		
クラス間分散 / エントロピー		0.5	0.75	1.0	0.5	0.75	1
0.5	葉の数	1			0.5	0.5	0.5
	木の深さ	0.6					
0.75	葉の数	0.9	0.9		0.7	0.7	0.5
	木の深さ	0.5	0.5				
1	葉の数				2.3	1.6	1.1
	木の深さ				0.4	0.3	
1.25	葉の数				3.9	3.9	1.6
	木の深さ				0.5	0.5	
1.5	葉の数				7	7	1.4
	木の深さ				0.8	0.8	

３．３　考察

実験結果が示すように、決定木を生成する際にクラスタ分析も用いることで推定精度が向上する場合がある。そのようなデータ集合の特性を見出すために、推定精度とデータ集合の特徴との関係を分析した。データ集合の特徴として、クラス内分散、クラス間分散、クラス内分散・クラス間分散比、クラス内のデータ間の平均距離、クラス間の平均距離を用いた。これらの特徴と、従来手法よりも本手法において推定精度が向上した割合との相関係数を求めた結果、クラス間分散およびクラス間の平均距離では -0.61 の負の相関がみられた。クラス内のデータ間の平均距離では-0.45、クラス内分散では-0.43 と、やや弱い負の相関が見られたが、クラス内分散・クラス間分散比では-0.14 であり、相関は見られなかった。これらの結果より、本手法では同一クラスのデータがまとまっているデータ集合において有用であることがわかった。これは決定木手法やクラスタ分析手法一般に言えることであることから、さらなる解析を行う必要がある。

４．おわりに

本稿では、決定木学習の精度向上を図るために、決定木を生成する際に従来手法に加え、クラスタ分析手法を用いる方法を提案した。評価実験の結果、データ集合の特性によっては推定精度向上を図ることができた。推定精度が向上したデータ集合の特性を分析した結果、明確な特徴は見いだせなかった。今後は、生成された決定木にさらに解析を加え、クラスタ分析の効果を明らかにするとともに、データ集合の特性に基づく閾値自動設定、クラスタ分析手法と決定木学習手法の効率的な切り替えなどのアルゴリズムの改善を行う。また、ランダム生成データや実データを用いた評価実験を行う必要がある。

参考文献

[1]Quinlan, J. R. 1986. “Induction of Decision Trees”, Machine Learning, Vol.1, No.1,pp.81-106(1986)

[2] Shekhar R.Gaddam et al,”K-Means+ID3:A Novel Method for Supervised Anomaly Detection by Cascading K-Means Clustering and ID3 Decision Tree Learning Methods”,Knowledge and Data Engineering,pp.345-354,2007

[3]寺邊ら,”相関ルールにもとづく属性生成手法”,人工知能学会誌,15 巻,1 号,pp.187-197 (2000)

[4]金　明哲,”R によるデータサイエンス”,森北出版株式会社,2007

[5]石井ら,”わかりやすいパターン認識”, オーム社, 1998

[6]MacQueen, J. B. (1967). ″Some Methods for classification and Analysis of Multivariate Observations″. 1. Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability. University of California Press. pp. 281-297. 12992. Retrieved 2009-04-07.