Available online at www.sciencedirect.com
ScienceDirect
ELSEVIER
December 2009, 16(6): 113–120
www.sciencedirect.com/science/journal/10058885
The Journal of China Universities of Posts and Telecommunications
www.buptjournal.cn/xben

New heuristic method for data discretization based on rough set theory

ZHAO Jun (✉), ZHOU Ying-hua

Institute of Computer Science and Technology, Chongqing University of Posts and Telecommunications, Chongqing 400065, China

Abstract

Data discretization contributes much to the induction of classification rules or trees by machine learning methods. The rough set theory is a valid tool for discretizing continuous information systems. Herein, a new method is proposed to improve those typical rough set based heuristic algorithms for data discretization, by utilizing decision information to reduce the scales of candidate cuts, and by more reasonably measuring cut significance with a new conception of cut selection probability. Simulations demonstrate that compared with other typical discretization algorithms based on the rough set theory, the proposed method is more capable and valid to discretize continuous information systems. It can effectively improve the predictive accuracies of information systems while still conceptually keeping their consistency.

Keywords data discretization, rough set theory, cut, cut significance, selection probability

1 Introduction

With the rapid development of information technologies and their wide applications, information systems with nominal, discrete, and/or continuous attributes are in exponential explosion. Nominal data can be easily transformed into discrete format without loss of information. Thus, these two data formats practically are not strictly distinguished and both of them are usually grouped into discrete type. From the perspective of knowledge induction, discrete and continuous types of data make significant differences in learning classification rules or trees, and discrete values usually outperform continuous ones in pentiful ways. Compared with continuous information systems, discrete systems generally are more capable of clustering instances and hence more robust to data noise; discrete attributes are closer to a knowledge level representation and hence easier to be understood, explained and applied; the induced knowledge from a discrete information system generally is shorter, more compact and accurate. Moreover, it is well known that some machine learning algorithms, like CN2 [1] and AQ [2], can handle discrete systems only. Though some other algorithms such as Naïve-Bayes Classifier [3] and C4.5 [4] etc, can directly deal with both discrete and continuous data, they usually perform much better over discrete information systems [5]. Consequently, data discretization, that is, a process of converting continuous values into a finite number of discrete ones, can significantly extend the borders of plentiful machine learning algorithms and greatly improve their performances, and thus is quite often necessary in practice.

Conceptually, data discretization partitions the attribute space of an information system into regions by a set of cuts, and makes all instances in a region with the same value vector. A discretized result can be evaluated in three dimensions [6]:

1) Simplicity: a discretization would be better if the attribute space is partitionized more roughly.

2) Consistency: a discretization would be better if less system inconsistency is increased.

3) Accuracy: a discretization would be better if more predictive accuracy is improved. Obviously, a practical discretization cannot score the highest in all dimensions, because there is always a tradeoff between simplicity and consistency imposed by the capability limit of data representation. The rough set theory [7] is one of the most

Received date: 21-10-2008
Corresponding author: ZHAO Jun, E-mail: zhaojun@cqupt.edu.cn
DOI: 10.1016/S1005-8885(08)60296-4

active and successful tools applied in data discretization. It takes 'instance discernibility' as a key conception. When processing an information system, the theory conceptually requires keeping its relation of instance discernibility. As a result, the degree of system consistency will be kept undisturbed in the sense of discernibility relation. Therefore, when discretizing an information system, the rough set theory can pursue the roughest partition of its attribute space. Meanwhile, it can conceptually remain its system consistency. This provides a natural way to deal with the tradeoff between simplicity and consistency of discretization.

Though a considerable amount of algorithms for data discretization have already been proposed based on the rough set theory，there is no statistically significant optimal one and more options are still solicited by various practical applications [8]. Herein, typical rough set based heuristic algorithms for data discretization are improved in two ways:

1) Decision information is introduced into the computation of candidate cuts. As a result, the scales of candidate cuts can be reduced remarkably and the real consumption of time and space can be saved greatly in the following steps;

2) A notion of cut selection probability is defined to measure cut significance more reasonably. Experiments indicate that the new measure is more suitable for computing result cuts for the ultimate discretization.

The remainder of this article is structured as follows. Sect. 2 gives some basic ideas about the rough set theory and data discretization. Sect. 3 proposes a new algorithm based on the typical heuristic discretization methods. Sect. 4 analyzes the complexities and tests the performances of the proposed algorithm. The article is concluded with a summarization in Sect. 5.

2 Rough set theory and data discretization

An information system is a tuple $\mathrm{DS}=(U,V,f,A\bigcup\{d\})$. Here, U, A and $\{d\}$ are sets of instances, condition attributes and a decision attribute, respectively, V is the value space of $A\bigcup\{d\}$, and f: $U\to A\bigcup\{d\}$ is the information function mapping instances to the attribute space.

Occasionally, DS is more exactly called a decision information system; $f(x,a)$ is also denoted by $a(x)$, where $x\in U$ and $a\in A\bigcup\{d\}$. For $x, y\in U$, x and y are discernible if $d(x)\neq d(y)\wedge\exists a\in A(a(x)\neq a(y))$. The mark dis($x$,$y$) denotes such a discernible instance couple. For $B\subseteq A\bigcup\{d\}$, B defines an indiscernibility relation ind(B) of U: $\mathrm{ind}(B)=\{(x,y)\,|\,(x,y)\in U\times U\wedge\forall b\in B[b(x)=b(y)]\}$. ind($B$) is obviously an equivalence relation of U. It determines a partition of U marked by $U/\mathrm{ind}(B)$. Particularly, $X\in U/\mathrm{ind}(A)$ is a so-called condition equivalence class, and $Y\in U/\mathrm{ind}(\{d\})$ is a so-called decision equivalence class.

For $a\in A$ of DS, if $V_a=[l_a,r_a)\subseteq\mathfrak{R}$ ($\mathfrak{R}$ is a set of real numbers) is its value domain, $P_a=\{[\,l_a,C_1^a),\,[C_1^a,C_2^a),...,\,[C_k^a,r_a)\}$, where $l_a<C_1^a<...<C_k^a<r_a$ and $V_a=[l_a,C_1^a)\bigcup\,[C_1^a,C_2^a)\bigcup...\bigcup[C_k^a,r_a)$, is a partition of V_a, and $C_a=\{C_i^a\,|\,i\in[1,k]\}$ is the set of cuts of a. A partition $P=\{P_a\,|\,a\in A\}$ of A is uniquely defined by sets of cuts $C=\{C_a\,|\,a\in A\}$. Obviously, P converts DS into a discrete system DS^P, and DS^P is the so-called discretization of DS. DS^P is consistent with DS if $\forall x$, $y\in U(\mathrm{dis}(x,y)$ in $\mathrm{DS}\to\mathrm{dis}(x,y)$ in $\mathrm{DS}^P)$; a consistent DS^P is irreducible if $\forall P'\subset P$ ($\mathrm{DS}^{P'}$ is not consistent with DS); an irreducible DS^P is optimal if its cardinality is not bigger than that of any other irreducible partition of DS.

From those related conceptions, one can conclude that data discretization based on the rough set theory conceptually requires keeping the relation of instance discernibility. Therefore, the discretization process has to implicitly or explicitly utilize the decision information of information systems. Compared with unsupervised methods, the utilization of decision information makes it more possible to keep undisturbed the instance distribution in attribute space, and to get good results even in cases where the instance distribution is not uniform.

When partitioning a continuous attribute, a direct method such as equal-width or equal-frequency determines all its intervals simultaneously. However, a discretization method based on the rough set theory usually works in an incremental way, that is, it begins with a simple discretization and then passes through an improvement process. Additionally, a criterion is usually necessary to stop the discretizing process.

3 A new heuristic method for data discretization based on the rough set theory

It has been shown that the optimal discretization is NP hard [9]. This fact clearly claims the importance of developing effective heuristic algorithms to figure out suboptimal discretization. Given an information system, it is typically discretized in three sequential steps:

1) To compute a set of candidate cuts.

2) To find out a subset, which is the so-called result subset