Development and Design of General Data Mining System

Chen Baowen[1,a]

[1]Department of College of Computer Science& Software Engineering, Shenzhen University, Shenzhen, China, 518060

[a]Bchenszsndd@163.com

Keywords: Data Mining; Optimal Design; discretization

Abstract. In this paper, we focus on top-down discretization methods and propose a new method for supervised discretization based on class-feature correlation by defining a class-feature contingency factor. The proposed method takes into consideration the distribution of all samples to generate an ideal discretization scheme. The method maintains a high interdependence between the target class and the discretized attribute, and avoids overfitting. Empirical evaluation of seven discretization algorithms on UCI real datasets show that the novel algorithm can yield a better discretization scheme that improves the accuracy of decision tree classification. As to the execution time of discretization and the number of generated rules, our approach also achieves promising results.

Introduction

The goal of discretization is to find a set of cut points to partition the range into a small number of intervals that have good class coherence, which is usually measured by an evaluation function. In addition to the maximization of interdependence between class labels and attribute values, an ideal discretization method should have a secondary goal to minimize the number of intervals without significant loss of class-attribute mutual dependence. Discretization is usually performed prior to the learning process and it can be broken into two tasks. The first task is to find the number of discrete intervals. Only a few discretization algorithms perform this; often, the user must specify the number of intervals or provide a heuristic rule. The second task is to find the width, or the boundaries, of the intervals given the range of values of a continuous attribute. In general, the algorithm for choosing landmarks can be either top-down, which starts with an empty list of landmarks and splits intervals, or bottom-up, which starts with the complete list of all the values as landmarks and merges intervals. In both cases there is a stopping criterion, which specifies when to stop the discretization process. Researchers in the machine learning community have introduced many discretization algorithms. Most of these algorithms perform an iterative greedy heuristic search in the space of candidate discretizations, using different types of scoring functions for evaluating a discretization.

Proposed Discretization Method

A. Rough Set Theory

Rough set theory was first proposed by Pawlak in 1980s. Now rough set is widely used in many aspects such as data mining, data analysis, knowledge discovery in database, text indexing, and so on.

In an information system, an appreciative space can be represented as the four-tuple $S=(U,A,V,F)$, where $U=\{x_1,x_2,\cdots,x_n\}$ is the universe that denotes a finite and non-empty set. $A=\{a_1,a_2,\cdots,a_m\}$ denotes a finite set of attributes. V_a denotes the domain of the attribute a. The attribute set $A=\{a_1,a_2,\cdots,a_m\}$ can be composed of decision attribute D and the condition attributes C. $V=\bigcup_{a\in A}V_a$ & $F:U\times A\to V$ is a total function such that $F(x,a)\in V_a$ for each $a\in A$, $x\in U$, called information function.

Approximation is the core concept of rough set. Let S be an information system, X a

non-empty subset of U and $\varphi \neq P \subseteq A$. The P-lower approximation and the P-upper approximation of X in S are defined, respectively, by:

Definition 1 If two elements x, $y \in U$ and $P \subseteq A$, θ_P is equivalent relation on U, if $x\theta_P y \Leftrightarrow (\forall p \in P)(f_p(x) = f_p(y))$, we say that θ_P is indistinguishable.

Definition 2 Set $X \subseteq U$, $P \subseteq C$, P-lower approximation about X can represent:

$$P\_X = \{x \in U \mid [x]_P \subseteq X\} \quad (1)$$

Where $[x]_P$ expresses equal kind of element set under the equivalent relation P.

Definition 3 Set $X \subseteq U$, $P \subseteq C$, P-upper approximation about X can represent:

$$P^- X = \{x \in U \mid [x]_P \bigcap X \neq \varphi\} \quad (2)$$

Definition 4 Set U denote the universe, P and Q are two equivalent relations bunches on U, Let us define $POS_P(Q)$ as follows:

$$POS_P(Q) = \bigcup_{X \in U/Q} P\_(X) \quad (3)$$

Definition 5 Set $P \subseteq C$, P-approximate precision (level of consistency) under partition $\{Y_1, Y_2, \cdots, Y_k\}$ is defined:

$$Lc_P = \sum_{i=1}^{k} card(P\_Y_i) / card(U) \quad (4)$$

where, $card(\,)$ expresses the cardinal number of set and Lc_P reflects the correct degree of decision classification.

B. Discretization Method

In this section, we propose propose a new method for supervised discretization based on class-feature correlation by defining a class-feature contingency factor, named CF(2). The proposed method takes into consideration the distribution of all samples to generate an ideal discretization scheme. The method maintains a high interdependence between the target class and the discretized attribute, and avoids overfitting.

Definition 7 Set X to be random variable which has limited value, $P_i = P\{X = x_i\}$, $i = 1, 2, ..., n$, then entropy of X is defined: $H(X) = \sum_{i=0}^{n} p_i \cdot \log_a \frac{1}{p_i}$

where logarithm bottom a can be any positive number, but generally take 2, when $p_i = 0$, $p_i \cdot \log_a \frac{1}{p_i}$.

The entropy defined above, is called the Shannon entropy generally, which was first proposed in 1948 by American Engineer C.E. Shannon, and afterward people improved it several times, but it did not have the great change. Therefore it has been used until now. But the Shannon entropy materially is only suitable for the limited situation. Continual random variable is not suitable. So, real value Attributes must be discretized, then we can use the Shannon entropy to carry on the information processing. At present, under the supervised form, correlative algorithms of discretization method based on information and entropy have been more influence, and they had been extensive researched. Mutual information theory had already been used in the discretization standard very early [5].

Algorithm

Algorithm 1

1 Input: Dataset with i continuous attribute, M examples and S target classes;
2 Begin
3 For each continuous attribute Ai
4 Find the maximum dn and the minimum d0 values of Ai;
5 Form a set of all distinct values of A in ascending order;
6 Initialize all possible interval boundaries B with the minimum and maximum

7 Calculate the midpoints of all the adjacent pairs in the set;
8 Set the initial discretization scheme as D: {[d0,dn]}and Globalcf(2) = 0;
9 Initialize k = 1;
10 For each inner boundary B which is not already in scheme D,
11 Add it into D;
12 Calculate the corresponding cf(2) value;
13 Pick up the scheme D' with the highest cf(2) value;
14 If cf(2) > Globalcf(2) or k < S then
15 Replace D with D';
16 Globalcf(2) = cf(2);
17 k = k + 1;
18 Goto Line 10;
18 Else
19 D' = D;
20 End If
21 Output the Discretization scheme D' with k intervals for continuous attribute Ai;
22 End

Experiments and Results

In order to evaluate our proposed algorithm in a real-world situation, 13 data sets are selected from the UC Irvine machine learning data repository [25] with numeric features and varying data sizes. The data are fully.

TABLE 1. The Summary of Data Sets

Data Sets	Continuous attributes attributes attributes	Discrete attributes attributes	Number of Classes	Examples
iris	4	0	3	150
auto	5	2	3	392
breast	9	0	2	683
Ionosphere	34	0	2	351
pima	8	0	2	768
glass	9	0	7	214
wine	13	0	3	178
machine	7	0	8	209
heart	5	8	2	296
sonar	60	0	2	208
vehicle	18	0	4	846
vowel	10	3	6	990
bupa	6	0	2	345

consistency or correct (inconsistency rate is zero), and the data contain real-life information from the medical and scientific fields which had been used previously in testing pattern recognition and machine learning methods. A summary of data sets can be found in Table1.

We compare our proposed method with the following algorithms for performance evaluation.

CADD: a popular top-down method;

CAIM: an outstanding top-down method;

MDLP: entropy-based method using the minimum description length principle;

EQW: a typical unsupervised top-down method.

Among the five discretization algorithms, EQW requires the user to specify in advance some parameters of discretization. For EQW, the number of intervals is set to 10. MDLP and our method have an automatic stopping rule and does not require any parameter setting.

In the following experiments, each data set is quantified respectively by the eight algorithms mentioned above. The 5-fold cross-validation test method is applied to all data sets. Each data set is divided into five parts, among which four parts are used as the training sets and one as the testing set. The experiments are repeated many times. The final predictive accuracy is taken as the average predictive accuracy value.

Conclusions

We proposed a static, incremental, supervised and bottom-up quantization algorithm in this paper, which presents a new quantization criterion and a heuristic algorithm. In order to estimate the effect of generated quantization schemes on the performance of the classification algorithm, empirical evaluation of existing quantization algorithms on UCI real data sets shows that our proposed method generates a better quantization scheme.

No matter uses what kind of merged standard to be able to have the influence to other attributes, but we hoped that the effect will achieve minimum. Moreover, the size of intervals has greater influence to merge, and this also is the question which the next step of work should take.

References

[1]H. Liu, F. Hussain, C. L. Tan, and M. Dash, "Discretization: an enabling technique," Journal of Data Mining and Knowledge Discovery, vol. 6, no. 4, pp. 393–423, 2002.

[2]J. Dougherty, R. Kohavi, and M. Sahami, "Supervised and unsupervised discretization of continuous feature," Machine learning: Proc. 12th Intl Conf., pp. 194–202, 1995.

[3]R. Kerber, "Chimerge: Discretization of numeric attributes," Proceedings Ninth National Conference on Artificial Intelligence, AAAI Press, pp. 123–128, 1992.

[4]Y. Geng, J. Chen, K. Pahlavan, Motion detection using RF signals for the first responder in emergency operations: A PHASER project, 2013 IEEE 24nd International Symposium on Personal Indoor and Mobile Radio Communications (PIMRC), London,Britain Sep. 2013

[5]Y. Geng, J. He, K. Pahlavan, Modeling the Effect of Human Body on TOA Based Indoor Human Tracking[J], International Journal of Wireless Information Networks 20(4), 306-317