A Improved Method of Discretization of Continuous Attributes

Shunling Chen¹, Ling Tang², Weijun Liu³, Yonghong Li³

¹Sichuan University of Science & Engineering Dept. of Science Engineering Zigong, Sichuan 643000, P. R. China
²Sichuan University of Science & Engineering Computer Application Research of SUSE Zigong, Sichuan 643000, P. R. China
³Sichuan University of Science & Engineering Dept. of Electronic Engineering Zigong, Sichuan 643000, P. R. China
csl_1220@163.com, tara_hq@yahoo.com.cn, weijun@suse.edu.cn, liyonghong652@163.com

Abstract

Discretization of continuous attributes is one of the important steps in preprocessing of data analysis. In this paper, a new method of supervised discretization of continuous attributes based on entropy and hierarchical clustering guiding by level of consistency of decision table is introduced. This method makes use of the concept of the level of consistency of decision table in Rough Sets. According to the level of consistency of the produced decision table, the number of hierarchical cluster is adjusted dynamically in the first step. And then in the second step, we merge adjacent region based on entropy without damaging the level of consistency. Experiments show that this method is feasible.

Keywords: Discretization, Clustering, Level of consistency, Entropy

1. Introduction

In processing of intelligence information, usually, the description of things is possibly qualitative explanation or expression of data; Characters expressed are hierarchical or out-of-order; Therefore, the attributes of database may be divided into two kinds: One is continuous (quantificational), they express some observable natures of objects, their values takes from some continuous region, such as, temperature, length and so on. The other is the discrete (qualitative), this kind of attributes are expressed with language or discrete values. Mostly, a database has contained both continuous attributes and discrete attributes. The continuous attributes discretization is a very important question in KDD. There are three kinds of discretization: based on whether the discretization has used category mark, the discretization may divide into the supervise discretization and the non-supervise discretization[4]; Based on whether the discretization time is different, the discretization may divide into the global discretization[1-2] and the local discretization; Based on whether consider the mutual influence between the attributes, may divide
into the static discretization[3] and dynamic discretization[5-6]. In view of the continuous attribute discretization, domestic and foreign scholars have proposed many methods, including the equidistant division method, the equal-frequent division method, the statistical method, the information entropy method and the clustering method, and so on. All of these methods divide the constitution space of attribute into finite number regions by way of the selected break point, and label each region with mark.

In this paper we propose an improved method of discretization of continuous attribute. The algorithm is divided into two parts. First we cluster data using the method of hierarchical clustering, and adjust dynamically the number of cluster according to the level of consistency of decision table, then form the initial division of each continuous attribute. Second merge adjacent region based on entropy, without affecting the level of consistency of decision table, it reduces the number of discrete regions.

2. Description of Discretization Question

A information system of database can be expressed by a quadruple \( S = \langle U, A, V, f \rangle \), where \( U \) is a finite set of object example, namely universe; \( A \) is attribute set, \( A = V \cup C \), where \( C \) is condition attribute set, \( D \) is decision attribute set; \( V = \bigcup V_a \), where \( V_a \) is range of attribute \( a \in A \); \( f \) is information function, it assigns the value of each example \( x \) of \( U \).

\( V_a = [la, Ra] \) is the rang of each continuous attribute \( a \in A \), and \((a, m)\) is break point of value \( c \) of \( a \). Then a classify \( P_a \) of \( V_a \) is defined by a random break point set \( \{(a, m_1), (a, m_2), \ldots, (a, m_n)\} \); and \( S' = \langle U, C \cup D, V, f \rangle \), where \( C' = \{a' : a'(x) = 1\} \), if \( a(x) \in [m_i, m_{i+1}] \). The original information system is substituted by a new system, after discretization.

2 The Definition of the Level of Consistency of Rough Set

We assume \( S = \langle U, R, V, f \rangle \) is a system of decision table, where universe \( U = \{x_1, x_2, \ldots, x_n\} \), and \( R = C \cup D \) is attribute set, \( C = \{c_1, c_2, \ldots, c_m\} \) is the condition attribute set, \( C \) and \( c_k (k = 1, 2, \ldots, m) \) are continuous attributes, \( D \) is the decision attribute set, \( V = \bigcup V_a \) is the set of value, where \( V_a \) expresses the range of \( a \in R \), and \( f : U \times R \rightarrow V \) expresses an information function, it is the set of value of attribute, it appoints each value of object of \( U \).

\( L_C \), the level of consistency of decision table, is defined by:

\[
L_C = \frac{\text{POS}_C(D)}{|U|} = \frac{\sum_{X \subseteq U \mid D} |C_\_X|}{|U|}
\]

Where \( \text{POS}_C(D) \) is the positive region of \( C \) of \( D \). \( \text{POS}_C(D) \) denotes object set of \( U \) that can be divided correctly, base on the classify \( U \mid C \), and added in the equivalence class of \( D \). And \( C_\_X \) is lower approximate set of \( X \).
3. The Measure of Entropy of Discretization of Data

Let $S$ is a set of $s$ samples, and suppose the attribute of decision is the attribute of class label with $m$ different values, and the different classes $C_i (i = 1, 2, \cdots, m)$ have been defined too. And $s_i$ is the number of sample of $C_i$, the expected information or the entropy which the classification of sample needs is given by the equation:

$$I(s_1, s_2, \ldots, s_m) = \sum_{i=1}^{m} p_i \log_2 p_i$$

Where $p_i$ is the probability of random sample belongs to $C_i$, and estimated with $s_i / s$.

Supposes condition attribute $A$ is a attribute of continuous value, according to expert's prior knowledge, $A$ contains $v$ hierarchies of concept. Therefore, there must be $v - 1$ thresholds for $A$, and $S$ is divided into $v$ subsets $S_j$, where $S_j \in \{S_1, S_2, \cdots, S_v\}$. If the maximal value of $A$ is $a_{\text{max}}$, the minimum value of $A$ is $a_{\text{min}}$, then the thresholds are $a_1, a_2, \cdots, a_{v-1}$ respectively, and the $v - 1$ thresholds divide interval $[a_{\text{min}}, a_{\text{max}}]$ into $[a_1, a_2, \cdots, a_{v-1}, a_{\text{max}}]$, where $a(x)$ is the value of sample $A$ of subset $S_j$ it should satisfies $a_{j-1} < a(x) \leq a_j$.

Supposes $s_{ij}$ is the number of sample of subset $S_j$ of $C_i$, after the concept hierarchies of $A$, the entropy of classification of sample is given by:

$$E(A) = \frac{\sum_{j=1}^{v} s_{ij} + s_{2j} + \cdots + s_{mj}}{s} I(s_{ij}, s_{2j}, \ldots, s_{mj})$$

Where $\left(\sum_{j=1}^{v} s_{ij} + s_{2j} + \cdots + s_{mj}\right) / s$ is the weight of $j$th subset. Obviously, the less entropy value is, the higher rate of classification base on the division of subset of $A$. According to the above equation, for a given subset $S_j$, such that

$$I(s_{ij}, s_{2j}, \ldots, s_{mj}) = \sum_{i=1}^{m} p_{ij} \log_2 p_{ij}$$

Where $p_{ij} = s_{ij} / |S_j|$ is the probability of $S_j$ belongs to $C_i$.

4. Hierarchical Cluster

When sample set $\{x_1, x_2, x_3, \cdots, x_N\}$ is known, we need to obtain counts $C$ of final cluster, the steps of hierarchical method are:

1) Let $R_j = \{x_j\}$, for all $j \in I, I = \{1, 2, 3, \cdots, N\}$;
2) Find out $R_i$ and $R_k$ in $\{R_j : j \in I\}$, they satisfy

$$\Delta(R_i, R_k) = \min \{\Delta(R_j, R_i)\}$$

Generally,

$$\Delta(R_i, R_k) = \min_{x \in R_j, y \in R_k} \{D(x, y)\}$$


\[ \text{D}(x, y) = |x - y| \quad \text{when sample are one-dimensional}. \]

3) Merge \( R_i \) and \( R_k \), then deletes \( R_i \);
4) Delete \( i \) from the index set \( I \), if the cardinal number of \( I \) is equal to \( C \), the method is terminated, otherwise go to 2).

5. **Improved Discretization Method of Continuous Attribute**

The method of discretization mainly divides into two parts. First, the initial division of the continuous attribute is carried on through hierarchical clustering. In the course of clustering, the level of consistency of decision table is computed after the division of cluster, so that the number of cluster is adjusted dynamically by the level of consistency. The second part’s main goal is to reduce the number of discrete region. Above all, we merge adjacent regions of zero entropy of each attribute, because this action doesn’t affect the level of consistency; If the level of consistency is still bigger than the primitive level, then gradually merge a pair of neighboring regions of the minimum entropy, until the level is equal to the primitive level, this process is simultaneously carrying on to all attributes. Consider influence of noise and disturbance, after the discretization, the level of consistency of decision table may be slightly less than that before discretization, then supposes the primitive level of consistency of decision table is added a very small number \( \delta \) on the post-discretization level of consistency of decision table, namely \( L_c = L_c^* + \delta \). The detail is as follows:

First step:

a) Assigns the decision system \( S = (U, A, V) \), \( A = C \cup D \), and supposes the number of condition attribute of \( C \) is \( n \), it means that the condition attribute is \( n \) dimensional, and appoints the number of initial classification of each condition attribute \( m_d = 2, d = 1,2, \cdots, n \); \( L_c^* \) is the acceptable level of consistency;

b) Let \( d = 1 \) at first;

c) With the method of hierarchical cluster, the division of \( d \) th dimension of condition attribute is the \( m_d \) kind;

d) Takes average of two boundaries of adjacent class as dividing value;

e) \( d = d + 1 \), returns to c), until \( d = n \);

f) Compute \( L_c \), if \( L_c \) is greater than or equal to \( L_c^* \), then go to second step, otherwise, let \( m_d = m_d + 1 \), for all \( d = 1,2, \cdots, n \), and go to b);

Second step:

g) Calculate the entropy of all adjacent regions of each attribute with the definition of entropy, and merge these adjacent regions of zero entropy.

h) If the level of consistency \( L_c - L_c^* \leq \delta \), then go to i), otherwise, compute entropy, and merge the adjacent regions of minimum entropy until \( L_c - L_c^* \leq \delta \);

i) Coding the value of discretization of attribute with digital 0,1,2,........

6. **Conclusion**

This paper proposes an improved method of discretization of continuous attribute. First carry on hierarchical clustering guiding by level of consistency of decision table, forms the initial division of each continuous attribute. Second merges adjacent ranges based on the entropy without affecting the level of consistency of the decision table. The method has been used in the data preprocessing of our research project of provincial educational committee, it works well. The practice has proved the validity of this algorithm.
7. Acknowledgment

This work was supported by the Scientific Research Fund of Sichuan Provincial Education Department (07ZB145)(2007R003).

References


