A Novel Discretizer for Knowledge Discovery Approaches Based on Rough Sets

Qingxiang Wu[1,2,3], Jianyong Cai[1], Girijesh Prasad[2]
TM McGinnity[2], David Bell[3], and Jiwen Guan[3]

[1] School of Physics and OptoElectronic Technology, Fujian Normal University
Fujian, Fuzhou, 350007, China
{qxwu, cjy}@fjnu.edu.cn
[2] School of Computing and Intelligent Systems, University of Ulster
Magee Campus, Londonderry, BT48 7JL, N.Ireland, UK
{Q.Wu, G.Prasad, TM.McGinnity}@ulster.ac.uk
[3] School of Computer Science, Queens University, Belfast, UK
{Q.Wu, DA.Bell, J.Guan}@qub.ac.uk

Abstract. Knowledge discovery approaches based on rough sets have successful application in machine learning and data mining. As these approaches are good at dealing with discrete values, a discretizer is required when the approaches are applied to continuous attributes. In this paper, a novel adaptive discretizer based on a statistical distribution index is proposed to preprocess continuous valued attributes in an instance information system, so that the knowledge discovery approaches based on rough sets can reach a high decision accuracy. The experimental results on benchmark data sets show that the proposed discretizer is able to improve the decision accuracy.

Keywords: Knowledge discovery, rough sets, continuous attribute discretization, decision-making, data preparation.

1 Introduction

Based on rough set theory, knowledge discovery, machine learning and data mining approaches [1,2] have been developed. For example, the multi-knowledge approach [3,4] is based on multiple reducts from rough set theory. Multi-knowledge approach can extract more useful knowledge from a training set so that a high decision accuracy can be reached. Because this approach prefers dealing with discrete data, a transformation from continuous values to discrete values is required. This is done using a continuous attribute discretizer. Two classes of discretizers (unsupervised and supervised discretizers) have been proposed in [5,6,7]. In this paper a new adaptive discretizer is proposed to solve the data type transformation problem in approaches based on rough sets. In this new discretizer, a distributional index is defined and applied to determine the splitting point within an interval. Based on the index decrement, the discretizer can adaptively discretize any continuous attribute without involvement of users. The discretizer

G. Wang et al. (Eds.): RSKT 2006, LNAI 4062, pp. 241–246, 2006.

can share statistical information with the multi-knowledge approaches and the Bayes classifier. The discretizer can also be applied to other machine learning approaches for discretization of continuous attributes. In Sect. 2, a statistical distribution is introduced. In Sect. 3, a algorithm for discretization is proposed. Experimental results and analysis are given in Sect. 4. Sect. 5 concludes the paper.

2 Statistical Distribution

2.1 Instance Information System

Following the notation in [2,4,8,12], let $I =< U, A \cup D >$ represent a instance information system, where $U = \{u_1, u_2, , u_i, , u_n\}$ is a finite non-empty set, called an instance space or universe, and where u_i is called an instance in U. $A = a_1, a_2, a_3, , a_i, , a_m$, also a finite non-empty set, is a set of attributes of the instances, where a_i is an attribute of a given instance. D is a non-empty set of decision attributes, and $A \cup D = \emptyset$. For every $a \in A$ there is a domain, represented by V_a, and there is a mapping $a(u) : U \rightarrow V_a$ from U to the domain V_a , where $a(u)$ represents the value of attribute a of instance u and is a value in the set V_a. For a given universe U, a domain of attributes is as follows.

$$V_a = a(U) = a(u) : u \in U for\ a \in A. \tag{1}$$

The domain of a decision attribute is represented by

$$V_d = d(U) = d(u) : u \in U for\ d \in D. \tag{2}$$

2.2 Value Number Distribution

In order to obtain a statistical table, a set of distribution numbers are defined as follows. Suppose that there is an instance information system $I =< U, A \cup D >$. Let N_{d_k,a_i,v_x} represent the number of instances with decision value d_k and attribute value $v_x \in V_{ai}$ for attribute a_i.

$$N_{d_k,a_i,v_x} = |u : d(u) = d_k\ and\ a(u) = v_x\ for\ all\ u \in U|. \tag{3}$$

Let N_{a_i,v_x} represent the number of instances with attribute value $v_x \in V_{ai}$ for attribute a_i.

$$N_{a_i,v_x} = |u : a(u) = v_x\ for\ all\ u \in U|. \tag{4}$$

2.3 Definition of Distributional Index

Based on principles of entropy of information [10,11], we construct a distributional index. Let $v_{st} \rightarrow v_{en}$ represent an interval of attribute value from value v_{st} to v_{en} and $N_{d_{main},a_i,v_{st}\rightarrow v_{en}}$ represent the number of instances that satisfies

$$N_{d_{main},a_i,v_{st}\rightarrow v_{en}} = \underset{d \in V_d}{max}(N_{d,a_i,v_{st}\rightarrow v_{en}}). \tag{5}$$

The distributional index is defined as follows.

$$E(v_{st} \rightarrow v_{en}) = \frac{-N_{d_{main},a_i,v_{st} \rightarrow v_{en}}}{|U|} log_n(\frac{N_{d_{main},a_i,v_{st} \rightarrow v_{en}}}{N_{a_i,v_{st} \rightarrow v_{en}}}). \quad (6)$$

where $|U|$ is the total number of instances in the instances information system, and n is the number of decision values. If $v_{st} \rightarrow v_{en}$ covers whole range of attribute values, $N_{a_i,v_{st} \rightarrow v_{en}} = |U|$. Suppose that all the values within this interval support one decision, i.e. $N_{d_{main},a_i,v_{st} \rightarrow v_{en}} = N_{a_i,v_{st} \rightarrow v_{en}}$. Therefore, we have the minimum of $E(v_{st} \rightarrow v_{en}) = 0$. If the $N_{d_k,a_i,v_{st} \rightarrow v_{en}}$ is an uniform distribution over the decision space, the maximum of $E(v_{st} \rightarrow v_{en})$ is equal to $N_{a_i,v_{st} \rightarrow v_{en}}/|U|$. This number decreases as more intervals are split.

3 Algorithm of Discretization

Based on the definition of the distributional index, a very simple algorithm is proposed to discretize continuous attributes. In order to discretize a continuous attribute, the number of intervals and the borders of intervals have to be determined. Let v_{border} represent the value of splitting point. The best splitting point can be found using following expression.

$$v_{border} = \underset{v_{bd} \in v_{st} \rightarrow v_{en}}{arg\ min}\ (E(v_{st} \rightarrow v_{bd}) + E(v_{bd} \rightarrow v_{en})). \quad (7)$$

According to the property of distribution index, the distribution index always becomes smaller when a interval is split into two intervals. Suppose that interval $v_{st} \rightarrow v_{en}$ is split into two intervals $v_{st} \rightarrow v_{bd}$ and $v_{bd} \rightarrow v_{en}$. The index decrement is defined as

$$\Delta E_{v_{st} \rightarrow v_{en}}(v_{bd}) = \{E(v_{st} \rightarrow v_{en}) - [E(v_{st} \rightarrow v_{bd}) + E(v_{bd} \rightarrow v_{en})]\}. \quad (8)$$

Based on this definition the splitting point can be rewritten as follows.

$$v_{border} = \underset{v_{bd} \in v_{st} \rightarrow v_{en}}{arg\ max}\ \Delta E_{v_{st} \rightarrow v_{en}}(v_{bd}). \quad (9)$$

For example, row 1 to 3 in Fig. 1 show a number distribution of Attribute 2 in the Wine data set. Applying Eq. 9 to this attribute, two intervals are obtained by splitting at the border v_{32} with maximal ΔE as shown in row 4 in Fig. 1. Applying Eq. 9 to the new intervals, the maximal decrement of the index can be obtained for splitting each interval. These new intervals and their the maximal decrements are put into a candidate list. The interval with largest maximal decrement in the candidate list is selected to split further. This splitting procedure is repeated until index decrement is zero for all the intervals or the desired number of intervals is reached. This is very different from the existing discretization approaches [5,6,7]. Row 4 to 7 in Fig. 1 show the discretization procedure of Attribute 2 in the Wine data set. Each row shows the curve of ΔE vs splitting point within the selected interval. The circle indicates the the splitting point with the maximal decrement.

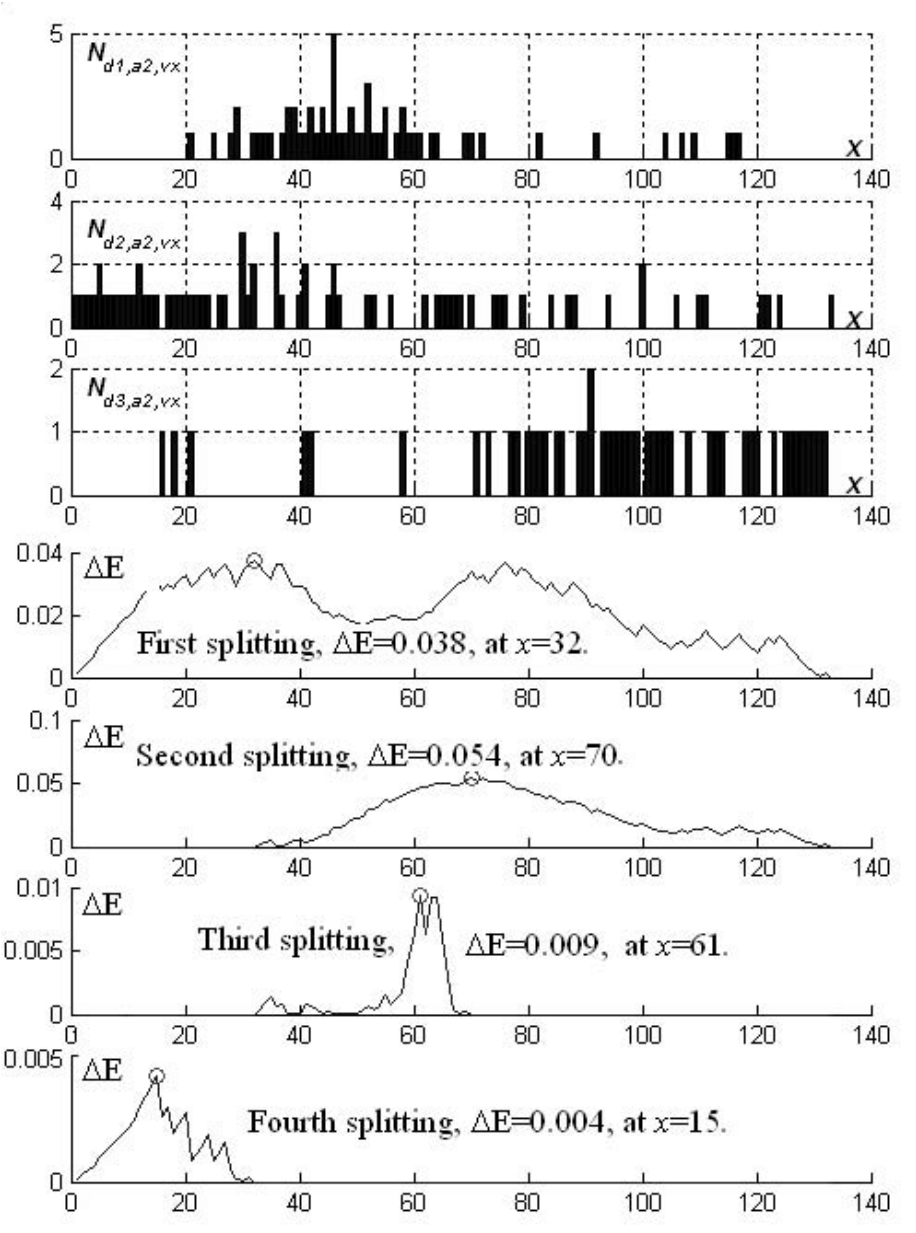

Fig. 1. Procedure of Discretization

4 Experimental Results

A set of 13 benchmark data sets from the UCI Machine Learning Repository [9] was applied to test both multi-knowledge approaches with the discretizer and

without the discretizer. The decision accuracies under the ten-fold cross validation standard are given in Table 1. Column 'No' lists decision accuracies for multi-knowledge approach without the discretizer. Column 'Dp' lists decision accuracies for multi-knowledge approach with the discretizer. In order to compare with an unsupervised discretizer, column '5e' lists decision accuracies for multi-knowledge approach with a 5-identical-interval discretizer. It can be seen that multi-knowledge approach with the adaptive discretizer improved decision accuracies for 13 data sets. The average accuracy over 13 data sets is better than multi-knowledge approaches without the adaptive discretizer and with a 5-identical-interval discretizer. Column C-type Attributes gives the number of continuous attributes contained in corresponding data set. The names with '*' indicate that some attribute values are missing in the data set.

Table 1. Comparison Results for New Discretizer An: Attribute Number, Cn: Continuous Attributes, In: Instance Number, No: No-Discretizer, Dp: Using the Proposed Discretizer, 5e: Using 5-Equal Discretizer

Data	An	Cn	In	No	Dp	5e
Sonar	60	60	208	77.8	97.1	91.4
Horse-colic*	27	7	300	80.0	86.3	80.3
Ionosphere	34	34	351	90.6	93.7	92.6
Wine	13	13	178	98.9	99.4	97.8
Crx-data*	15	6	690	85.1	86.5	85.0
Heart	13	6	270	83.3	86.3	85.1
Hungarian*	13	6	294	85.4	85.4	84.0
SPECTF	44	44	80	73.8	98.8	92.5
Bupa	6	6	345	65.5	70.2	67.0
Iris-data	4	4	150	96.7	96.7	93.3
Ecoli	6	6	336	71.5	75.3	75.0
Anneal*	38	6	798	99.4	99.7	99.7
Bands*	39	20	540	77.8	79.6	76.5
Average				83.5	88.8	86.2

5 Conclusion

In this paper a new discretizier based on the distributional index is proposed. The minimum of the distributional index is applied to determine the border value for splitting an interval. The maximum of index decrement is applied to select the new intervals to split further. This discretizier has combined with both information entropy and statistical distributions so that quality of rules exacted from data sets can be improved after the discretization. Therefore, high decision accuracies can be obtained. As number distributions are also applied in the naive Bayes classifier and the multi-knowledge approaches [4,12], this discretizier can be combined with the naive Bayes classifier and the multi-knowledge approaches with very little increase of computational cost. The discretizer has been combined

with the multi-knowledge approach to making decision. The experimental results on 13 benchmark data sets show that the average accuracy has been improved. This discretizer can be combined with other machine learning approaches for further study.

References

1. Lin, T.Y. and Cercone, N., Eds.: *Rough Set and Data Mining.* Kluwer Academic Publishers, (1997).
2. Polkowski, L., Tsumoto, S., and Lin, T.Y., Eds.: *Rough Set Methods and Applications: New Developments in Knowledge Discovery in Information Systems.* Physica-Verlag, A Springer-Verlag Company, (2000).
3. Hu, X, Cecone, N., and Ziarko, W.: Generation of multiple knowledge from databases based on rough set theory. 1 (1997) 109-121.
4. Wu, Q.X., Bell, D.A, and McGinnity, T.M.: Multi-knowledge for Decision Making. *International Journal of Knowledge and Information Systems.* Springer-Verlag, 2 (2005) 246 - 266.
5. Dougherty,J., Kohavi, R., and Sahami, M.: Supervised and Unsupervised Discretization of Continuous Features. In: Proceedings of International Conference on Machine Learning, (1995) 194-202.
6. Wu,X.: A Bayesian Discretizer for Real-Valued Attributes. *The Computer J.* 8 (1996) 688-691.
7. Kurgan, L.A., and Cios, K.J.: CAIM Discretization Algorithm. *IEEE Transaction on Knowledge and Data Engineering.* 2 (2004) 145-153.
8. Pawlak, Z.: *Rough sets: theoretical aspects data analysis.* Kluwer Academic Publishers, Dordrecht, (1991).
9. Blake, C.L. and Merz, C.J.: UCI Repository of Machine Learning Databases, *http://www.ics.uci.edu/ mlearn/MLRepository.html.* UC Irvine, Dept. Information and Computer Science, (Download in 2003).
10. Quinlan,J.R.: Induction of Decision Trees. *Machine Learning.* 1 (1986) 81 - 106.
11. Mitchell, M.T.: *Machine Learning. McGraw Hill Co-published* by the MIT Press Companies, Inc. (1997).
12. Wu, Q.X., and Bell,D.A.: Multi-Knowledge Extraction and Application. In: Wang, G.Y., Liu, Q., Yao, Y.Y., and Skowron, A., Eds. *Rough Sets, Fuzzy Sets, Data Mining, and Granular Computing (RFDGrC03).* LNAI 2639, Springer, Berlin, (2003) 274-279.