Advanced Science and Technology Letters
Vol.31 (MulGraB 2013), pp.261-270
http://dx.doi.org/10.14257/astl.2013.31.56

Research on RSIEDM Algorithm and the Application to Lightning Disaster Statistics

Shuonan Hou[1], Rongtao Hou[2], Xinming Shi[2], Jun Wang[2],
Chengshang Yuan[2], Jin Wang[2]

[1](College of Information Science and Engineering, Northeastern University, ShenYang 110819, Liaoning, China)

[2](School of Computer & Software, Nanjing University of Information Science & Technology, Nanjing 210044, Jiangsu, China.)

Abstract. This paper proposes a new method based on Rough set theory-information entropy- discernible matrix discretization. The method uses rough, information entropy and Discernible matrix that can be more reasonable and more accurately to continuously attribute discretization, and making created decision tree have better accuracy. In the application of optimization of lightning disaster statistics and evaluation result of lightning disaster, the algorithm which has obtained a better effect.

Key words: decision tree, discretization, rough set, information entropy, discernible matrix

1 Introduction

Decision tree has been widely used in classification, machine learning, and knowledge discovery. The decision tree method originated from concept learning system (CLS) which has the drawback that it can not deal with big problems. In 1986, J.Ross Quinlan proposed a new method of Interative-Dichotomizer 3(ID3). C5.0 is an improved algorithm for ID3. ID3 assumes that the properties are discrete values; however, many properties are continuous in practical situation. These properties are discretized by C5.0 in the way of range dividing, which reduces the classification accuracy.

Rough set that proposed by Pawlak is a new mathematical theory for dealing with imprecise, incomplete and incompatible knowledge. It has obtained widespread application in various aspects. Many valuable result of the study based on the rough set model have been got by researchers. Nguyen proposed a method that based on rough set and Boolean logical. This method can find all the possible discrete data set, but its algorithm complexity is exponential, which reduces the possibility of application in practical situation [5]. Some researchers proposed several improved greedy algorithm to overcome this drawback [4, 6]. These algorithms are local optimization search algorithm which based on breakpoint separability of instance. Some researchers used genetic algorithm, which belong to global search algorithm, to search the best discretization set breakpoints [7, 8]. Some researchers proposed a kind

ISSN: 2287-1233 ASTL

of algorithm which based on the importance of attributes [9]. Some researchers proposed a discretization method with the combination of polynomial hyper surface and support vector machine [10] . Some researchers proposed a clouds-mode based discretization method [11]. Some researchers Introduced fuzziness into discretization [12], some researchers discussed Discretization of information granularity [13]. Some researchers proposed the attributes-clustering based discretization method [14].

To improve the accuracy of discretization method, we proposed an Attribute discretization method based on RSIEDM (Rough Set Theory-Information Entropy and Discernible Matrix) to overcome the drawback in the procedure of C5.0. RSIEDM use Theory-Information Entropy to discretize the continuous attributes and then reduce attributes through discernibility matrix.

2 Decision Tree Classification Algorithm

Classification is a greedy algorithm. When a training set was given, decision tree partitioning feature attribute space by recursion, then pick the highest information gain rate of properties as training attribute of current node to guarantee the simplest decision tree [1].

2.1 Create Decision Tree

One of the attribute training conditions was necessarily selected to partition data set into several smaller subset in every recursion of creating decision tree. C5.0 algorithm test attributes by information gain ratio, so the growing procedure of decision tree will end at a necessary constraint condition, such as when all the nodes are divided into their class, and all the records have the same attribute. Also other standard can be used to terminate decision tree growth process in advance, such as pre pruning technology [2]. Assume that S is a training sample set. X which contains n attributes divides S into n subsets S_1, S_2... S_n. Assume the count of samples in S is |S|, $freq(C,S)$ is the number of sample which belongs to C_i (i=1,2,...,N), the probability of a sample belongs to C_i is $\log_2(\text{freq}(C,S))/|S|$. The training set can be presented as formula (1). The $info(s)$ is the gross information content which is necessary for identify all the samples in S.

$$info(S) = \sum_{i=1}^{N} \frac{freq(C_i, S)}{|S|} \log_2\left(\frac{freq(C_i, S)}{|S|}\right) \quad (1)$$

After divide S into n subsets, the information entropy of each subset can be calculated. The value of expectation of S is shown in formula (2).

$$info_x(S) = \sum_{j=1}^{n} \frac{|S_i|}{S} info(S)_i \quad (2)$$

In order to measure the information of S which partitioned by X according to

attribute verification, the information gain standard gain(X) shown in formula (3) was used. The attribute with the highest information gain was chose for partition.

$$gain(X) = info(S) - info_x(S_i) \quad (3)$$

The method divided S into $S_1, S_2 \ldots S_n$ based on the different values of X, the potential information produced by these subsets can be presented by formula (4).

$$Split\ Info_X(X) = -\sum_{i=1}^{n} \frac{|S_i|}{|S|} \log_2\left(\frac{|S_i|}{|S|}\right) \quad (4)$$

Hence, the information gain ratio of S partitioned by X is shown in formula (5).

$$Gain\ ratio = \frac{\Delta\ info}{Split\ Info} \quad (5)$$

C5.0 algorithm regards the attribute with highest information gain ratio as training attribute, and prunes the original decision tree by post-pruning algorithm. Pruning a decision tree usually means that to replace one or numbers of tree by a leaf node, and then the node with highest probability of occurrence is regarded as a class.

2.2 Discretize the Continuous Valued Attribute

C5.0 algorithm discretized the continuous valued attribute better than ID3. However, it is a heavy computation algorithm. If the amount of these continuous valued attributes is huge, the computation amount will be more. According to the lightning disaster database that collected by meteorological department, the attributes in lightning disaster database are shown in Table 1.

Table 1. description of attribute of lightning disaster data

Attribute description	**Domain**
OT	Datetime
WB	0-1
LN	0-1
RN	0-1
OME	0-1

ND	0-**x**
NI	0-**x**
DEL	0-**x**
IEL	0-**x**
NC	0-**x**

The top 5 discrete attributes are occurrence time (OT), wind (WD), lighting (LN), raining (RN), and other meteorological elements(OME). The next 5 attributes are continuous valued. They are number of death (ND), number of injury (NI), direct economic loss (DEL), indirect economic loss (IEL), and number of casualties (NC). 0-1 stands for the value if the accident happened. 0-**x** stands for the continuous value. A means normal, B means serious, C means critical. The numbers of samples in class A, B, and C respectively are 1579 (81%), 317 (16.2%) and 24 (2.7%). The total number is 1947. 5 attributes among them are continuous valued attribute which are the primary attributes that need to be discretized. According to these continuous valued attributes, the time efficiency and accuracy of C5.0 is reduced.

3 Discretized Research based on Rough Set Information Entropy Discernible Matrix

Many discretization methods had been proposed, and different method may lead to different result. All these methods should follow such principles: the dimension is as little as possible after discretizated and most information of attribute should be maintained after discretizated.

Discretization often affects the process and the final results of the algorithm. Hence, an appropriate discretized method should be used. The RSIEDM reduce the redundant attributes and attribute values on the basis of considering all the attributes to improve accuracy of discretization.

3.1 Algorithm of Initial Breakpoints Set

To determine the initial breakpoint set is the basis to solve the problem of attribute discretization. On the premise of guarantee the distinguishing relationship of decision table, how to make the base of the initial set breakpoint is as small as possible to have very important sense to the subsequent work. It can not only reduce the amount of calculation of the initial set breakpoints, and can reduce the computation time and space.

Definition 1. Ordered sequence of attribute values. Decision table DT={u, R, V, f}, R=CU{d}. Continuous attributes a .∈ CAssume that Ia=[V_a , v_{aT}] is value range

$V_a <$

of a, if $^{B}\ V_a < V_{a^2} < \ldots < V_{a^T}$, then sequence S={ ^{B} V_a, V_a^1, V_a^2, ..., V_{aT}} is called ordered sequence of attribute values. V_{aB} is the greatest lower bound of s, called B(S). ^{T} V_a is the least upper bound of s, called T(S).

According to Definition 1, let $S[m]=min(T(S_1),T(S_2))$ and $S[n]=max(B(S_1),B(S_2))$ the relationship of ordered sequences S_1 and S_2 may have four cases below, $S[m]< S[n]$, $S[m]= S[n]$, $S[m]> S[n]$ and $S_2 \square S_1$

The values of attribute ***a*** are divided into several subsets on the basis of decision condition. These subsets make up their corresponding ordered sequences. The initial breakpoint sets are empty set. If S_i and S_j (i<j) in ordered sequences are in the cases 1 and 2, then max(B(Si),B(Sj)) should be added to P. Continue checking the remained attributes, decide the sequence number **m** and **n** of max(B(Si),B(Sj)) and min(T(Si),T(Sj)), then add S[m] and S[n] into P. For element S[k] in S between S[m] and S[n], if S[k-1] and S[k] are in the Si or Sj in the same time, then ignore S[k], or add S[k] into P. Finally, check whether there is any sequence of attribute values, if no sequence exists, then set P is a breakpoint set of **a**.

3.2 The Information Entropy of Rough Set Attributes Discretization Method

Information entropy is a measure of the uncertainty in information system attribute. Method based on information entropy uses information provided by classes. The following steps are information entropy based discretization method.

Usually, the smaller the information entropy is, which indicates that individual decision attribute values are dominant, the smaller the degree of chaos. Especially when if and only if the attribute values of instance of X are the same, the information entropy equals zero. This nature guarantees the breakpoint reduction algorithm does not change the compatible degree of decision table.

Hence, the information entropy of X according to breakpoint ^{a} c_j is formula (6).

$$H^{X}(c^{a}_{j}) = \frac{|X_b|}{|X|} H(X_b) + \frac{|X_t|}{|X|} H(X_t) \qquad (6)$$

For concluding information entropy, assume that $L=\{Y_1, Y_2, \ldots, Y_m\}$ is a equivalence class which divided from decision table by set Q, then after a new breakpoint c□Q be added, the new information entropy is as shown in formula (7).

$$H\ c\ L = H^{} + H^{} + + H^{}$$

$$(\ ,\ )\qquad {}^{1}(c)\qquad {}^{2}(c)\ \ldots\qquad {}^{m}(c) \qquad (7)$$

$H(c, L)$ Becomes smaller, which indicates that the decision attributes of new equivalence class become single after added the breakpoint. Hence, $H(c, L)$ indicates the importance of breakpoint **c**.

Assume Q is reduction breakpoint set, L is equivalence set that partitioned by breakpoint set Q, H is the information entropy of decision table. In practical situation, not all the possible cut points should be took into consideration.

3.3 Attribute Reduction of Discernibility Matrix

After the continuous values of attributes are divided into discrete space by breakpoints, two drawbacks will be encountered. 1) If there are too many decision classes in decision table, the particle number of decision table after discretization will be high, in subsequent planning or classification process the superiority of the discrete data will not be manifested. 2) The discretization method will easily produce some isolated range for the decision table with noise. It seriously affected the quality of data mining model formed subsequently. To avoid these drawbacks, discernibility matrix could be used to optimize discretization procedure.

Definition 2. Redundant attribute. For $a_i \in B(B \subseteq C)$, if $POS_B(D) = POS_{(B-\{a_i\})}(D)$, then a_i is a redundant attribute, which means it will not affect the B-lower approximation of D, or a_i is necessary.

Definition 3. Relative core and reduction of attribute. Assume $B \subseteq C$, if $POS_B(D) = POS_C(D)$, $\forall a_i \in B$, and $POS_{(B-\{a_i\})}(D) \neq POS_B(D)$, then B is a reduction of A. The set composed of all the decision attributes D in condition attribute set C is called core of C (relative core). It is also a set which composed of single element in discernibility matrix. Discernibility matrixs of C and its core are $n \cdot n$ symmetric matrices, $n = card(U)$, and any element a(x,y) in them is shown as follows:

$$c(x,y) = \{a \in C \mid f(x,a) \neq f(y,a), x \in [x]_{Ind(D)}, y \in [y]_{Ind(D)}, [x]_{Ind(D)} \neq [y]_{Ind(D)}\}$$

Core of C is a set composed of all the single element in the discernibility matrix.

$U/C \Rightarrow$		a_1	a_2	a_3	D
	$\{X_1, X_3, X_9\}$	2	1	3	1
	$\{X_2, X_7, X_{10}\}$	3	2	1	2
	$\{X_4\}$	2	2	3	2
	$\{X_5, X_8\}$	1	1	4	3
	$\{X_6\}$	1	1	2	3

Fig.1. Decision table

Figure 1 shows the basic set of C after attribute discretized. Then the discernibility matrix shown in Figure 2 could be deduced. X1, X2 ...Xn in the matrix stand for each discretized lightning record. We can conclude from from Figure 2 that {a2} is the core of C, its discernibility matrix is $f_C(D) = a_2(a_1 + a_3) = a_1a_2 + a_2a_3$. $f_C(D)$ shows there are two reduction of decision table, they are{a1,a2} {a2, a3}.

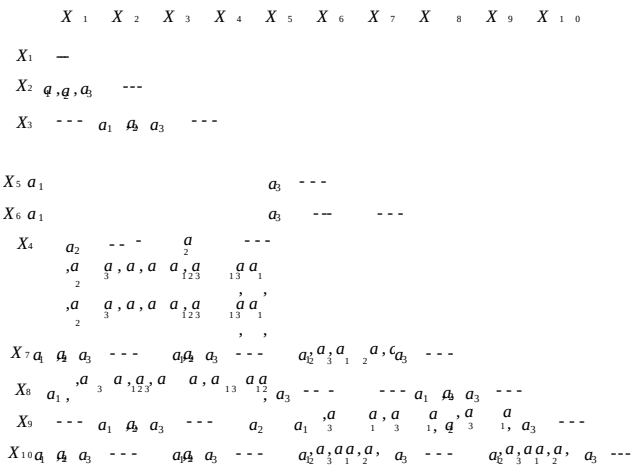

Fig. 2. Discernibility matrix of C

Further reduction of attributes could be processed after attributes reduction. The basic sets of system can be gained even if a part of unnecessary values are wiped out. To find the procedure of finding core and reduction of attribute values is the same as the procedure of finding the core and reduction of attributes.

After above procedures, the breakpoints set P^a of continuous value $a \in C$ can be calculated. After dealing with the other attributes with the same method, we can get all the breakpoints set P. At this point, the procedure of discretized method is ended.

4 Experimental Verification

To verify the effectiveness of the algorithm and compare the performance of the

algorithm with other discrete algorithm, we tested data sets (Iris, Breast, Wine, Sonar) in database of UCI and lightning data (LzData). The continuous valued attributes in lightning database are direct economic loss, indirect economic loss, number of death, number of injury, and total of injury. The information of data used in experiments was shown in Table 2.

Table 2. Description of information of data set

Name of dataset	Number of continuous attribute	Number of class	Number of instance
Iris	4	3	150
Breast	9	2	683
Wine	9	7	214
Sonar	60	2	208
LzData	5	3	600

In contrast, we proposed RSIEDM in this paper and Ext_Chi2 for discretization. Ext_Chi2 is an improved algorithm of Chi2 which is supervised and based on statistics theory. In the experiment, we used Support vector machine technology to classify the discrete data. The method of One to many(1-V-r) was chosen for classification, C-SVC was chose for model type, RBF was chosen for kennel function, [1, 100] is the search range for penalty factor, [0.05, 0.5] is search range for kernel function parameters γ. The 80% of data set was chosen for the training set; the remained 20% of data set was chose for test set. After classification, the prediction accuracy was shown in Table 3.

Table 3. classification results of SVM(1-V-r)

Comparative indicators	Name of dataset	Discretized method	
		Ext_Chi2	RSIEDM
Prediction accuracy(%)	Iris	100.0	93.8
	Breast	65.2	74.6
	Wine	95.6	97.3
	Sonar	60.6	76.2
	LzData	65.2	73.7

The experiment result shows that the proposed algorithm has certain advantages compared with other algorithms when dealing with a large amount of dataset. Ext_chi2 gain 100% accuracy when dealing with Iris dataset due to its bottom-up algorithm. It can always get good effect when dealing with small sample data. However, for the dataset with a large number of instances, the series of Chi2 Algorithm not only processing slowly, but also gain worse effect than the top-down algorithm.

5 Conclusions

From the results of the experiment, the average prediction accuracy is higher than C5.0 algorithm. Furthermore, our proposed algorithm shows high performance when dealing with large amount of data. It is a useful discretization method. The future research work includes improving the algorithm, combining rough sets theory and other heuristic algorithms, further improving the efficiency of solving the problem of large-scale data.

Acknowledgements. Our research was supported by the National Natural Science Foundation of China(No. 61373064), University Graduate and Undergraduate Student Scientific Research Innovation Projects CXLX12_0515 and 201310300016Z in Jiangsu Province. And was supported by the teaching innovation project 13JY001 of Nanjing University of Information Science & Technology.

References

1. Matthew S Sullivan,Martin J Jones,David C Lee,et al.A com-parison of predictive methods in extinction risk studies: Con-trasts and decision trees [J]. Biodiversity and Conservation, 2006,15(6):1977-1991.
2. WANG XiZhao YANG ChenXiao. Merging-Branches Impact on Decision Tree Induction [J]. Chinese Journal of Computers e,2007, 30(8):1251-1258.
3. LIU Peng, YAO Zheng, YIN Junjie. Improved decision tree of C4.5 [J]. Journal of Tsinghua University(Science and Technology),2006,46(S1):996-1001.
4. Wang Guo2Yin. The Theory of Rough Set and Knowledge Ac2quisition. Xi′an: Xi′an Jiaotong University Press, 2001(in Chinese) .
5. NGUYENH S, SKOWRON A. Quantization of real values attributes,rough set and Boolean reasoning approaches[C] //Proc of the 2nd JointAnnualConference on Information Science, Wrightsville Beach:[s. n.], 1995: 34-37.
6. DAI Jian-hua, LIYuan-xiang·Study on discretization based on rough set theory[C] //Proc of the 1st International Conference onMachine Learning and Cybernetics. 2002: 1371-1373.
7. CHEN Cai-yun,LI Zhi-guo,QIAO Sheng-yong,etal.Study on discre-tization in rough setbased on genetic algorithm[C] //Proc of the 2nd International Conference on Machine Learning and Cybernetics.2003: 1430-1434v
8. HUANG Jin-jie, LI Shi-yong·A GA-based approach to rough data model[C] //Proc of the 5thWorld Congress on IntelligentControland Automation. 2004: 1880-1884v
9. HOU Lijuan,Wnag GuoYi, Nie Neng. Discretization in Rough Set Theory[J]·Computer Science, 2000,27(12): 89-94.
10. HE Yaqun, HU Shousong. A New Method for Continuous Value Attribute Discretization in Rough Set Theory. Journal of Nanjing University of Aeronautics & Astronautics,

2003,35(3).
11. Li Xingsheng. A New Method Based On Cloud Model For Discretization of Continuous Attribute in Rough Sets[J]. PR&AI, 2003,16(3): 33-38.
12. ROY A, PAL SK·Fuzzy discretization of feature space for a rough set classier[J]·Pattern Recognition Letters, 2003,24(6): 895-902.
13. WANG Li-hong, ZHANG Shu-cu,i FAN Hu,iet al.The information granulation in discretization[C] //Proc of the 2nd InternationalCon-ference onMachine Learning and Cybernetics. 2003: 2620-2623.
14. LIMeng-xin,WU Cheng-dong,HAN Zhong-hua,etal.A hierarchical clusteringmethod for attribute discretization in rough set theory[C] //Proc of the 3rd InternationalConference onMachineLearning and Cy-bernetics. 2004: 3650-3654.

