Entropy-based Algorithm for Discretization

Lamia AbedNoor Muhammed
University of Al-Qadisiyah

Entropy-based Algorithm for Discretization

by
Lamia Al-sawwaff
Computer Science Dep.
College of Computer and Mathematic Sciences
University of Al-Qadissiya
e-mail: lamia@qadissuni.edu.iq

Abstract

Discretization is a common process used in data mining applications that transforms quantitative data into qualitative data. Different methods have been proposed in order to achieve this process. The stand stone in the discretization algorithm is to find potential cut-points which split continuous range values into nominal values. So the discretization methods vary according to how to find these cut-points. Entropy based method is one of discretization methods however using information entropy measure.

In this paper, the aim was to use the entropy-based method in the discretization with a proposed algorithm. This algorithm attempts to find suitable cut-points through new concepts. It is based on the entropy information method with statistical tool through several steps . The practical work was executed on experimented data that were downloaded from UCI repository data.

Keywords:- Discretization, Entropy, Split, Merge

1- Introduction

Discretization is a common process used in data mining applications that transforms quantitative data into qualitative data. Discretization is an essential task of data preprocessing not only because some learning methods do not handle continuous, but also for other important transformed in a set of intervals are more cognitively relevant for a human interpretation[1].

Many data mining systems work best with qualitative data, where the data values are discrete descriptive terms. However, lots of data are quantitative being represented by a numeric value rather a small number of descriptors. One way to apply existing qualitative systems to such quantitative data is to transform the data[2]. The goal of descritization is to find a set of cut points to partition the range of continuous values into 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 the 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[3].

Different discretization algorithms' are exist, but traditionally they can be can be divided into two categories:

1- Unsupervised (or class-blind) algorithms discretize attributes without taking into account respective class labels. The two representative algorithms are **equal-width** and **equal-frequency** discretization.

2- Supervised algorithms discretize attributes by taking into account the interdependence between class labels and the attribute values. The representative algorithms are:, **Information Entropy Minimization**, **Statistics-based algorithms**, **class-attribute interdependency algorithms**, and **clustering-based algorithms**[4].

While in recent years, extended supervised discretization algorithms have been emerged and known as "semi-supervised" discretization algorithms. These algorithm based on the concept; by reducing the information needed to execute supervised algorithms i.e. non-parametric semi-supervised discretization method that based on the MODL framework (“Minimal Optimized Description Length”)[**].

2-Discretization algorithm

Discretization algorithm is the set of steps that are required in order to transform continuouse values into another expression discrete values. It aims to find the cut-points. The term”cut-points refers to a real value within the range of continuous values that divides the range into two interval is less than or equal to the cutpoint and the other interval is greater than the cut-point[4].

Atypical discretization algorithm broadly consists of four steps:

(1) sorting the continuous values of the feature to be discretized, (2) evaluating a cut-point for splitting or adjacent intervals for merging, (3) according to some criterion, splitting or merging intervals of continuous value, and(4) finally stopping at some point.[3]

3- **Entropy Measure and Discretization**

Entropy is one of the most commonly used discretization measure in the literatures. Shannon defines entropy of a sample variable X as:

$$H(X) = -\sum_{x} p_x \log p_x \quad \text{.........} \ (1)$$

Where x represents a value of X and p_x its estimated probability of occurring. It is the average amount of information per event where information of an event as:

$$I(x) = -\log p_x \quad \text{.........} \ (2)$$

Information is high for lower probable events and low otherwise. Hence, entropy H is the highest when each event is equi-probable[5].

Discretization methods use entropy measures to evaluate candidate cutpoints. This means that an entropy-based method will use the class information entropy of candidate partitions to select boundaries for discretization. Class information entropy is a measure of purity and it measures the amount of information which would be needed to specify to which class an instance belongs[3].

Definition 1: Let T partition the set S of examples into the subset S1,…..Sn. Let there be k classes C1,……..Ck and let p(Ci, Sj) be the proportion of examples in Sj that have class i. The class entropy **Ent()** of a subset Sj is defined as[7]:

$$Ent(Sj)=\sum p(Ci,Sj) \log(p(Ci,Sj)) \quad \text{.........} \ (3)$$

It has been shown that optimal cut points for entropy minimization must be between examples of different classes.

The generalize classes to multiple attributes is defined as:

Definition 2: Let A is a set of attributes {A1,……..Am} that can be used in measured of classes entropy C11, ……. ,Cmi. Where Cmi is the state (i) of attribute (m). The class entropy of Sj is defined as:

$$Ent(Sj)= \sum\sum p(Cmi,Sj) \log(p(Cmi,Sj)) \quad \text{.........} \ (4)$$

3- Proposed descritization algorithm

The proposed descritization algorithm in this paper is based on using information entropy methods with iterative steps. Also, the algorithm using statistical measure that is standard deviation, i.e. this measure support the selection of cut-point. The proposed algorithm based on multiple stages. First, primary discretization for continuous attribute can be executed using equal-width, second step is the splitting for candidate subsets(intervals), third step is merging for specific intervals. In each step there is some criteria would be used. The proposed algorithm is shown in figure(1).

```
1- Prepare data for processing
2- Sorting the interest continuous data X={x1,……xn}
3- Specified a class attribute C={c1,..cm}
4- Discretize continuous attribute's values in primary discretization
   with subsets  X={S1,….Sk}
5- Compute information entropy Ent(S) according to class
   attribute C
6- Compute standard deviation of Ent
7- do while Ent(Sj)>threshold(s)
      split Ent(S) into two subsets
      Compute Ent(S) for new subsets
8- Do while no. of interval>threshold(i)
   -  Merge the subsets that can produced minimum total Ent and
      standard deviation
   -  Compute Ent(I) for new subsets
```

Figure(1) proposed algorithm

4- Experiment and Results

4-1 description of experiment data

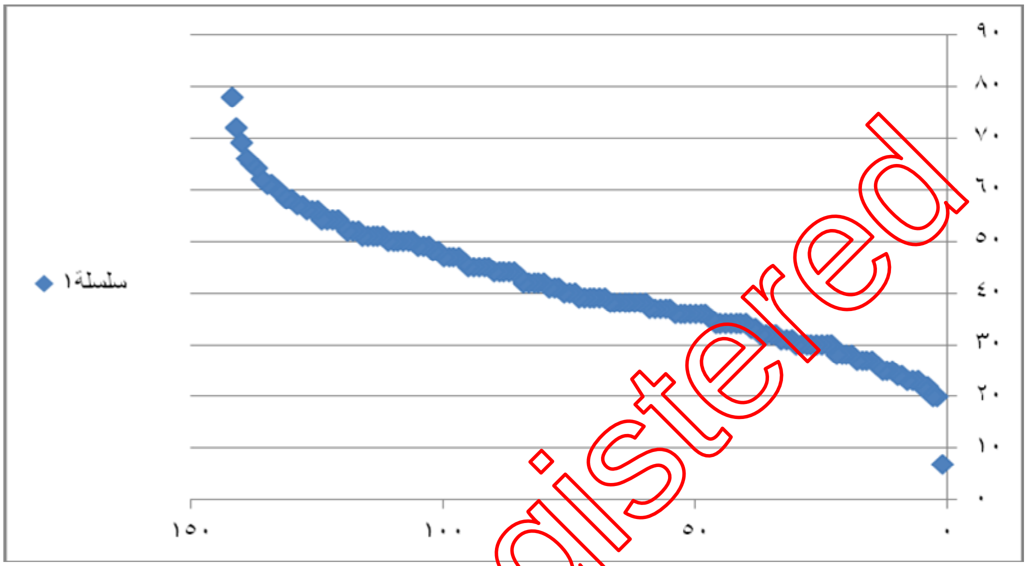
The practical work was accomplished using data set that are available in UCI repository for machine learning researches. The data set is "Hepatitis". It consists of 19 attributes(categorical, integer, real) with 155 instances.

Attribute Information:

1. Class: DIE, LIVE
2. AGE: 10, 20, 30, 40, 50, 60, 70, 80
3. SEX: male, female
4. STEROID: no, yes
5. ANTIVIRALS: no, yes
6. FATIGUE: no, yes
7. MALAISE: no, yes
8. ANOREXIA: no, yes
9. LIVER BIG: no, yes
10. LIVER FIRM: no, yes
11. SPLEEN PALPABLE: no, yes
12. SPIDERS: no, yes
13. ASCITES: no, yes
14. VARICES: no, yes
15. BILIRUBIN: 0.39, 0.80, 1.20, 2.00, 3.00, 4.00
16. ALK PHOSPHATE: 33, 80, 120, 160, 200, 250
17. SGOT: 13, 100, 200, 300, 400, 500,
18. ALBUMIN: 2.1, 3.0, 3.8, 4.5, 5.0, 6.0
19. PROTIME: 10, 20, 30, 40, 50, 60, 70, 80, 90
20. HISTOLOGY: no, yes

4-2 Results

The proposed algorithm was experimented on "Hepatitis" data set. The interest continuous attribute that to be discretized was "AGE". The distributed of examples in the data set according to attribute "AGE" is shown in figure(2).



Figure(2) distribution of examples according to "AGE" continuous values

All the nominal attributes in data set were candidate to be class attribute and experimented in primary discretization step of proposed algorithm. Each candidate class attribute was discretized with **equal-width**. The widths of interval(subsets) were (30,20,15,10,7). The information entropy for each subset was computed and then total entropy and standard deviation were computed as shown in Table(1). The difference between the candidate attributes in produced result is obvious. Three attributes have best results (SEX, ASCITES, VARICES). These attributes would be candidate to the next step of proposed algorithm that is splitting step.

The resulted data from equal-width partition using (7 width) that concerned with "SEX" attribute are (9 partitions). Then resulted data would be passed to next step of descritization that is "splitting". But the condition of splitting was not occurred according to splitting threshold(s) that was chosed(0.7) so there is no splitting and the data were passed to the merging step. In the merging step , the partitions that were merged are 8 and 9, 1 and 2, then 7 merged with 8 and 9. So the result partitions are 6 partitions. Each partition is addressed by one interval of "SEX" attribute as shown in table(2).

interval no.	Interval range	entropy value
1	-26	0.05898433
2	27-33	0.04847932
3	34-40	0.05439528
4	41-47	0.04785283
5	48-54	0.04785283
6	55-	0.05578864

Table(2) partitianing attribute **AGE** into interval according to

attribute SEX class attribute value

The resulted data from equal-width partition using (7 width) that concerned with " ASCITES " attribute are (9 partitions). Then resulted data would be passed to next step of descritization that is "splitting". Only partition (5) was candidate for splitting because its entropy is exceed threshold(s). After splitting, then the merging step would be executed. The partitions that were merged are 1 and 2,9 and 10, then 4and 5, 6 and. The results are 6 partitions as shown in table(3).

interval no.	range	entropy value
1	-26	0.040434003
2	27-33	0.048479321
3	34-41	0.055553853
4	42-47	0.079202133
5	48-54	0.047852827
6	55-	0.055788643

Table(3) partitianing attribute **AGE** into interval according to

attribute **ASCITES** class attribute value

At last attribute "VARICES" was candidate to be the a class attribute and its concerned data are obtained from equal-width partition using (15 width). Then resulted data were passed to splitting. Only partition (5) was candidate for splitting because its entropy is exceed threshold(s). After splitting, then the merging step would be executed. The partitions that were merged are 1 and 2,9 and 10, then 4and 5, 6 and. The results are 6 partitions as shown in table(4).

interval no.	range	entropy value
1	-28	0.028647625
2	29-36	0.069408734
3	37-39	0.044218153
4	40-43	0.036729102
5	44-47	0.064257689
6	48-51	0.052859411
7	52-	0.029567632

Table(4) partitianing attribute **AGE** into intervals according to attribute **VARICES** class attribute value

4-3 Discussion

1-The results that were obtained from primary step discretization reveal the difference between different attributes concerned with computing entropy values and so this step is useful in order to choose the attribute that candidate to be class attribute, therefore three with minimum entropy value would be chosen.

2- The practical work in this paper experiment random variable in measuring the width of intervals to be produced.

3- Using primary step can be terminated when the entropy value will be less, or the standard deviation is not improved through using different width(this case can be notice with **VARICES** attribute) as shown in table(1).

4- Using standard deviation with entropy value to guide the splitting or merging steps.

5- Conclusion

The execution of proposed algorithm reveled some facts:

1- The trained examples must have good distribution for continuous attribute that to be discretized in order to produce good result.

2- It is convenient to be considered the standard deviation of entropy value for the total partitions in order to forbidden the

3- Using primary discretization with "equal width" method is suitable for decrease the iterations of splitting

References

[1] Bondu, A. and Boulle, M. and Lemaire, V., Loiseau, S. and Duval, B., '
A Non-parametric Semi-supervised Discretization Method
ICDM '08 Proceedings of the 2008 Eighth IEEE International Conference on Data Mining, IEEE Computer Society Washington, DC, USA .

[2] Fayyad U., Irani K., Multi- Interval Discretization of Continuous-Valued Attributes for classification Learning, Proccedings of the thirteenth Int. Joint conference on Artificial Intelligence, Amsterdam, John Wiley & Sons, 1994, p. 284-432.

[3]Kotsiantis, S. and Kanellopoulos, D., 2006, “Discretization Techniques: A recent survey”, GESTS International Transactions on Computer Science and Engineering, vol.32(1), p.p. 47-58.

[4] Kurgan, L. A. and Cios, K. J., 2004, “CAIN Discretization Algorithm”, “IEEE Transactions on Knowledge and Data Engineering”, vol.16, No.2.

[5] Liu, H., and others, 2002, “ Discretization: An Enabling Technique”, “Data Mining and knowledge Discovery”, no.6, p.p.393-423, Kluwer Academic Publishers. Manufactured in Netherlands.

[6] Muhlenbach, F. and Rakotomalala, Ricco, 2005, “Discretization of Continuous Attributes”, Encyclopedia of Data Warehousing and Mining, John Wang(Ed) 2005, p.p.397-402.