Discretization of continuous attributes using ReliefF

Marko Robnik Šikonja
Igor Kononenko
University of Ljubljana
Faculty of Electrical Engineering and Computer Science
Tržaška 25,
61001 Ljubljana,
Slovenia
tel.: +386 61 1768-386, fax: +386 61 264-990
{Marko.Robnik, Igor.Kononenko}@fer.uni-lj.si

Abstract

Many existing learning algorithms expect the attributes to be discrete. Discretization of continuous attributes might be difficult task even for domain experts. We have tried the non-myopic heuristic measure ReliefF for discretization and compared it with well known dissimilarity measure and discretizations by experts. An extensive testing with several learning algorithms on six real world databases has shown that none of the discretizations has clear advantage over the others.

1 Introduction

Discretization divides values of the continuous attribute into a number of intervals. We represent each interval by a symbol and treat the attribute as discrete.

Many inductive learning algorithms expect continuous attributes to be discretized before learning; they use some discretization algorithm as preprocessing. Discretization is often considered a task for domain expert and used as background knowledge for specific problem, but often she is unavailable or unable to specify exact intervals. With the help of automatic procedure the task becomes easier.

Discretization of attributes can reduce the learning complexity and help to understand the dependence between attributes and the target concept. Also, in the constructive induction process, the discretized attributes are needed for most operators [13].

There are several methods we can use to discretize an attribute. We can divide them into two groups.

- Class blind techniques (equal width, uniform frequency, maximal marginal entropy) are easy to implement, but require user to specify the number of intervals. As they ignore class of examples when partitioning them into different intervals, the classification algorithm might no longer be able to separate examples of different classes. The essential information may be lost and consequently the method performs poorly in many situations [11].
- Class aware methods try to partition the examples into intervals that most closely correspond to the concept. They differ about the measure they use to estimate the closeness.

The heuristic measure we use is ReliefF [6] (an extension of Relief [4]) which is a powerful attribute estimator performing well on artificial as well as real world problems [8].

We have used ReliefF to guide the discretization process and compared the results with another discretization which uses attribute dissimilarity [9, 3]. We have tested both measures with several machine learning algorithms and databases.

2 Discretization algorithm

We have used the following greedy algorithm for discretization of attributes (similar algorithm was used by [3]):

```
BestDiscretization = {}
SetOfBoundaries = {}
repeat
  add the split which maximizes the heuristic
   measure to the SetOfBoundaries
  if SetOfBoundaries is best so far then
    BestDiscretization = SetOfBoundaries
until n times the heuristic is worse
      then in previous step
```

At each step the algorithm searches for the value which, when added to the current set of boundaries,

maximizes the heuristic estimate of the discretized attribute.

We compared performance of two heuristic estimates.

ReliefF [6] is an extension of Relief [4] which can deal with missing values and multivalued classes and also extends the power of estimation.

Intuitively, we can present ReliefF as estimating probabilities that two near learning examples from the same class have the same value of the attribute and that two near examples from different class have the same value of the attribute. The first probability testifies for the predicative power of attribute, while the second testifies against it. Two near examples are defined as having small distance in the space of all attributes. These two probabilities for different near surroundings are combined into a single estimate with range from -1 to 1. The larger the estimate the more significant the attribute.

For comparison to our novel approach we have used a well known measure of dissimilarity between attribute and class [9]. It is defined as:

$$D(A,C) = \frac{H(A|C) + H(C|A)}{H(A\&C)}, \qquad (1)$$

where $H(X)$ represents entropy of variable X. It is defined as

$$H(X) = -\sum_x p(x) \log p(x) \qquad (2)$$

where x runs over all possible values of X.

The measure takes into account the conditional entropy of both the dependent and the independent variable, and normalizes them with entropy of their co-occurrence.

3 Learning algorithms

The main advantage of ReliefF to other attribute estimators is its non-myopic behaviour. ReliefF can namely predict the quality of strongly dependent attributes. To observe if this feature helps in discretization of continuous attributes, we have selected three learning algorithms capable of some sort of constructive induction, and two without such capability.

All the algorithms except the Semi-naive Bayesian classifier are top-down decision tree builders equipped with pruning and handling of missing values similar to Assistant's [2]. We have reimplemented [12] LFC (Lookahead Feature Construction) [10], an algorithm who, at each node of the tree, joins attributes with conjunction, disjunction and negation and uses these constructs instead of original attributes to split the learning set. RCI (ReliefF in Constructive Induction) is similar to LFC, but it uses ReliefF instead of the posterior entropy to estimate the quality of attributes (constructs). It has an additional constructive operator (equality).

The naive Bayesian classifier presumes independence of attributes and therefore uses the naive Bayesian formula to predict classes of unseen examples. The Semi-Naive Bayesian classifier [5] is its extension: it tries to join certain values of attributes into new attributes, hoping to capture crucial dependencies. Formation of new attributes is a kind of constructive induction.

Figure 1 summarizes important features of learning algorithms.

Name	Heuristic	Construction
Assistant-I	entropy	no
LFC	entropy	yes
Assistant-R	ReliefF	no
RCI	ReliefF	yes
SN-Bayes	Bayes	yes

Figure 1: Some features of selected learning algorithms

4 Experimental scenario

We have undertaken an extensive testing of both discretization methods on all five learning algorithms on six real world and one artificial domain. Figure 2 gives short description of testing domains.

Domain	attr	cont	cl	ex	%
Breast cancer	10	4	2	288	80
Diabetes	8	8	2	768	65
Heart	13	7	2	270	56
Hepatitis	19	6	2	155	79
Iris	4	4	3	150	33
Rheumatology	32	21	6	355	66
XOR2-cont	12	12	2	512	50

Figure 2: Description of testing domains (attr=number of all attributes, cont=number of continuous attributes, cl=number of classes, ex=number of examples, %=majority class percentage)

All the real world domains are well known in machine learning community and can be acquired from various sources (eg. Irvin machine learning database repository at ftp://ftp.ics.uci.edu/pub/machine-learning-databases/). Each of the these databases

already has the discretization boundaries available. These boundaries were mostly suggested by domain experts. We have also tested the algorithms on these discretizations. Results on these discretizations can be considered an informative benchmark.

The artificial domain represents continuous version of order 2 parity (XOR) problem. Two attributes are informative while others are given random values.

Figure 3 illustrates this problem.

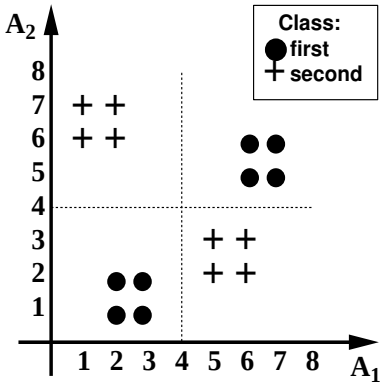

Figure 3: Continuous version of eXclusive OR. Ideal discretization would split both attributes at 4.

For each domain we have made 10 splits of the data into training and testing set consisting of 70% and 30% of examples respectively. For each split we have used the training set to perform discretization and learning, then we tested on testing set. The results are averaged over ten runs.

We compared discretizations and learning methods in terms of the prediction accuracy and the information score [7]. The later takes the prior probabilities into account and is defined as

$$Inf = \frac{\sum_{i=1}^{N} Inf_i}{N} \qquad (3)$$

$$Inf_i = \begin{cases} -\log_2 P(Cl_i) + \log_2 P'(Cl_i) & P'(Cl_i) \geq P(Cl_i) \\ \log_2 P(1-Cl_i) - \log_2 P'(1-Cl_i) & P'(Cl_i) < P(Cl_i) \end{cases}$$

where $P(C)$ is prior probability of correct class C, $P'(C)$ the probability predicted by classifier and N is the number of tested examples.

5 Results

As one can see from figures 4, 5, 6, 7 and 8, for real world problems there is no clear winning discretization with neither of learning algorithm.

Results for the artificial problem given in figure 9 clearly indicates advantage of discretization with ReliefF. Only this discretization has preferred the boundary 4 for both informative attributes. All the algorithms were able yo exploit this information, especially the ones with constructive induction.

Domain	Assistant-I			Assistant-R		
	g	d	R	g	d	R
Breast	78.0	76.0	72.7	77.4	76.5	75.1
Diab.	72.0	73.8	72.0	70.8	67.4	69.2
Heart	76.9	78.1	76.2	78.2	79.2	76.6
Hepa.	81.1	81.2	83.2	78.0	79.5	78.8
Iris	95.3	94.1	94.2	96.0	93.9	93.7
Rheum.	62.3	62.2	62.2	64.0	60.8	65.2

Figure 4: Classification accuracy (in %) for algorithms without construction for given (g), dissimilarity (d) and ReliefF's (R) discretization

Domain	Assistant-I			Assistant-R		
	g	d	R	g	d	R
Breast	-0.02	-0.00	-0.06	-0.01	-0.04	-0.04
Diab.	0.31	0.35	0.28	0.23	0.20	0.22
Heart	0.51	0.52	0.50	0.49	0.50	0.46
Hepa.	0.26	0.27	0.30	0.17	0.19	0.20
Iris	1.47	1.43	1.44	1.48	1.41	1.42
Rheu.	0.42	0.41	0.44	0.35	0.30	0.42

Figure 5: Information score (in bits) for algorithms without construction for given (g), dissimilarity (d) and ReliefF's (R) discretization

6 Conclusion

The results show that in spite of ReliefF's theoretical advantage both discretizations perform similarly on real world domains.

We also have to mention that ReliefF's discretization is computational more demanding.

The constructive induction algorithms did not show any advantage over the plain inductive learners. This leads as to the conclusion that the tested real world domains are too easy (the attributes are probably independent), so neither the ReliefF's discretization nor the constructive induction could show its real power.

Our conclusion in a nutshell is: if you suspect your domain has dependent attributes, try ReliefF's discretization, otherwise such heavy artillery is needless.

References

[1] Breiman L., Friedman L.H, Olshen R.A., Stone C.J.:

Domain	LFC			RCI		
	g	d	R	g	d	R
Breast	75.9	73.8	71.2	76.7	77.8	75.0
Diab.	69.0	73.8	69.5	70.5	72.0	72.0
Heart	77.3	78.0	78.4	73.1	77.7	76.6
Hepa.	80.6	82.7	80.6	78.5	76.6	80.1
Iris	94.2	94.6	94.9	95.2	90.2	92.6
Rheum.	62.4	57.8	63.2	64.6	65.5	63.9

Figure 6: Classification accuracy (in %) for algorithms with construction for given (g), dissimilarity (d) and ReliefF's (R) discretization

Domain	LFC			RCI		
	g	d	R	g	d	R
Breast	-0.01	-0.02	-0.06	-0.07	-0.02	-0.07
Diab.	0.27	0.35	0.28	0.23	0.29	0.28
Heart	0.53	0.54	0.55	0.42	0.48	0.45
Hepa.	0.25	0.31	0.25	0.18	0.14	0.23
Iris	1.46	1.46	1.46	1.47	1.36	1.39
Rheu.	0.42	0.31	0.47	0.33	0.32	0.36

Figure 7: Information score (in bits) for algorithms with construction for given (g), dissimilarity (d) and ReliefF's (R) discretization

Classification and regression trees. Wadsworth Inc., Belmont, California, 1984

[2] Cestnik, B., Kononenko, I., Bratko, I.: ASSISTANT 86: A Knowledge-Elicitation Tool for Sophisticated Users. In: Ivan Bratko, Nada Lavrač (eds.): *Progress in Machine Learning, Proceedings of EWSL 87.* Wilmslow, Sigma Press, 1987.

[3] Cestnik, B.: Informativity-Based Splitting of Numerical Attributes into Intervals. In Hamza, M.H.(ed):Expert Systems Theory & Applications, Proc. of the IASTED International Symposium,, Acta Press, 1989

[4] Kira K., Rendell L.: A practical approach to feature selection. In *Proceedings of the Machine Learning Conference, Aberdeen*, 1992, pp. 250-256

[5] Kononenko, I.: Semi-Naive Bayesian Classifier, *Proc. of EWSL 91*, Porto, Portugal, March 6-8, 1991, pp.206-213

[6] Kononenko, I.: Estimating attributes: analysis and extensions of RELIEF. *Proc. of European Conf. on Machine Learning*, Springer Verlag, 1994.

[7] Kononenko, I., Bratko, I.: Information based evaluation criterion for classifier's performance. *Machine Learning 6*, 1991, pp. 67-80

[8] Kononenko, I., Šimec, E.: Induction of decision trees using RELIEFF. In Kruse, R., Viertl, R., Della Riccia, G. (eds.): *CISM Lecture notes*, Springer Verlag, 1995 (in press)

[9] Mantaras R.L.: ID3 revisited: A distance based criterion for attribute selection. *Proc. Int. Symp. Methodologies for Intelligent Systems*, Charlotte, North Carolina, USA, October 1989

[10] Ragavan H., Rendell L.: Lookahead Feature Construction for Learning Hard Concepts. *Proceedings of the Tenth International Machine Learning Conference, 1993*, pp. 252-259

[11] Richeldi M., Rossotto M.: Class-Driven Statistical Discretization of Continuous Attributes. In Lavrač N., Wrobel S.: *Machine Learning: ECML-95*, Springer 1995

[12] Robnik, M.: Konstruktivna indukcija z odločitvenimi drevesi. Diplomska naloga. Univerza v Ljubljani, FER, Ljubljana, 1993

[13] Robnik, M.: Konstruktivna indukcija v strojnem učenju. *Elektrotehniški vestnik*, Vol. 62(1):43-49

Semi-naive Bayesian classifier						
Domain	accuracy			inf. score		
	g	d	R	g	d	R
Breast	77.9	79.2	78.1	0.06	0.07	0.07
Diab.	76.6	73.4	74.3	0.38	0.31	0.32
Heart	82.2	84.8	81.5	0.66	0.71	0.67
Hepa.	85.1	84.4	83.9	0.37	0.34	0.35
Iris	95.3	93.9	94.5	1.51	1.49	1.49
Rheum.	67.9	67.8	67.1	0.53	0.56	0.56

Figure 8: Classification accuracy (in %) and information score (in bits) for Semi-naive Bayesian classifier for given (g), dissimilarity (d) and ReliefF's (R) discretization

Continuous XOR problem				
Algorithm	accuracy		inf. score	
	d	R	d	R
Assistant-I	87.7	87.7	0.76	0.76
Assistant-R	90.5	93.5	0.71	0.81
LFC	96.7	96.7	0.93	0.93
RCI	90.8	100	0.77	1.00
SN-Bayes	87.4	93.2	0.78	0.87

Figure 9: Classification accuracy (in %) and information score (in bits) of all learning algorithms on continuous version of XOR problem, for dissimilarity (d) and ReliefF's (R) discretization