Discretization as a Means of of Enriching the Set of Features

Avi Rosenfeld, Ron Illuz, Dovid Gottesman, and Mark Last
Jerusalem College of Technology
Ben Gurion University of the Negev
rosenfa@jct.ac.il, mlast@bgu.ac.il

1 Extended Abstract

Discretization is a data preprocessing technique which transforms continuous attributes into discrete ones. This is accomplished by dividing every attribute A into m discrete intervals where $D = \{[d_0, d_1], (d_1, d_2], \ldots, (d_{m-1}, d_m]\}$ where d_0 is the minimal value, d_m is the maximal value and $d_i < d_{i+1}$ for $i = 0,1, ..., m\text{-}1$. The resulting values within D constitute a discretization scheme for attribute A and $P = \{d_1, d_2, ..., d_{m-1}\}$ is A's set of cut points. Traditionally, discretization has been used in place of the original values such that after preprocessing the data, D is used instead of A [Garcia and others, 2013]. We refer the reader to a recent survey [Garcia and others, 2013] for a detailed comparison of existing discretization algorithms and other algorithms.

This paper's first claim is that discretization should not necessarily be used to replace a dataset's original values, but instead to generate new features to augment the existing dataset. As such each discretized value D should be used in *addition* to the original value A. Using discretization in this fashion is to the best of our knowledge a completely novel idea. As such, we propose using datasets with both A and D as a method for further improving the performance of a classification algorithm for a given dataset.

This paper's second claim is that discretization algorithms should be developed with the purpose of adding new features to the original data. As support for this point, we present D-MIAT, an algorithm that *d*iscretizes data based on *M*inority *I*nteresting *A*ttribute *T*hresholds. D-MIAT focuses on identifying important features where only the minority of an attribute's values strongly point to one of the classes needing to be learned. We claim that at times it can be important to create discretized features with such indications. However, attribute selection approaches to date typically treat all values within a given attribute equally, and thus focus on the general importance of all values within a given attribute, or combinations of the full set of different attributes' values. As such, these approaches would typically not focus on cut points based on strong indications within only a small subset of values.

To support these claims we studied the prediction accuracy of 28 datasets within a previous discretization study [Cano *et al.*, 2016], the results of which are summarized in Figure 1. We considered the prediction accuracy of 5 different machine learning algorithms– Naive Bayes, K-nn, Adaboost, C4.5 and Logistical Regression, within 4 different types of

Discretization can Enhance Performance

Baseline	74.06	80.45	52.57	79.16	80.57	73.36
Baseline + MIAT	74.66	80.58	52.57	79.23	80.99	73.6
Ameva + Orig	75.53	79.51	52.74	78.39	81.27	73.49
CAIM + Orig	76	79.2	52.66	78.53	81.72	73.62
CACC + Orig	74.95	79.71	52.54	78.5	80.55	73.25
CHI + Orig	76	80.41	52.73	76.98	81.59	73.54
EF + Orig	74.78	77.16	52.59	76.77	81.14	72.49
EW + Orig	75.07	75.94	52.75	77.53	81.44	72.55
HDD + Orig	75.36	75.78	52.7	72.61	80.53	71.4
IDD + Both	75.5	78.82	52.71	77.79	81.26	73.22
IEM + Orig	75.75	80.91	52.57	78.94	82.1	74.05
ur-CAIM+Orig	75.63	80.07	52.79	78.62	81.58	73.74
Combination of MIAT and other Algorithms Works Best						
Baseline	74.06	80.45	52.57	79.16	80.57	73.36
CAIM + MIAT Orig	76.27	79.81	52.66	78.83	81.85	73.88
IEM + MIAT Orig	76.21	81.23	52.57	79.08	81.89	74.2
ur-CAIM+Orig +MIAT	75.92	80.53	52.76	79.2	82.36	74.15

Figure 1: Using Discretization to Enrich Datasets

datasets. First, we considered the accuracy of original dataset (A) without any discretization. We then compared this accuracy to the discretized datasets (D) from the 10 algorithms previously considered [Cano *et al.*, 2016], using the training and testing data within that paper. Surprisingly, we found that prediction accuracy on average *decreased* when only the discretized data was considered. The third dataset we considered was the combination of the first two datasets, appending the discretized values from each of the 10 discretization algorithms to the non-discretized data. We also created a dataset with the original data and D-MIAT alone. In addition, we created an 11th dataset with the original values (A) and the discretized values created by D-MIAT. Overall, the prediction accuracy of this third type of data outperformed that of the previous two types of data. We then proceeded to create a fourth dataset that combined the discretized values of D-MIAT and the three discretization algorithms with the best prediction performance. The combination of the original data, plus that of the "standard" discretization algorithms plus D-MIAT performed the best.

References

[Cano *et al.*, 2016] Alberto Cano, Dat T Nguyen, Sebastián Ventura, and Krzysztof J Cios. ur-caim: improved caim discretization for unbalanced and balanced data. *Soft Computing*, 20(1):173–188, 2016.

[Garcia and others, 2013] Salvador Garcia et al. A survey of discretization techniques: Taxonomy and empirical analysis in supervised learning. *IEEE Transactions on Knowledge and Data Engineering*, 25(4):734–750, 2013.