Optimal Bin Number for Equal Frequency Discretizations in Supervized Learning

MARC BOULLE
France Telecom R&D
2, Avenue Pierre Marzin
22300 Lannion – France
marc.boulle@francetelecom.com

Abstract. While real data often comes in mixed format, discrete and continuous, many supervised induction algorithms require discrete data. Although efficient supervised discretization methods are available, the unsupervised Equal Frequency discretization method is still widely used by the statistician both for data exploration and data preparation. In this paper, we propose an automatic method, based on a Bayesian approach, to optimize the number of bins for Equal Frequency discretizations in the context of supervised learning. We introduce a space of Equal Frequency discretization models and a prior distribution defined on this model space. This results in the definition of a Bayes optimal evaluation criterion for Equal Frequency discretizations. We then propose an optimal search algorithm whose run-time is super-linear in the sample size. Extensive comparative experiments demonstrate that the method works quite well in many cases.

Key words: Data Mining, Machine Learning, Discretization, Bayesianism, Data Analysis

1 Introduction

Discretization of continuous attributes is a problem that has been studied extensively in the past [6, 8, 10, 14, 21]. Many classification algorithms rely on discrete data and need to discretize continuous attributes, i.e. to slice their domain into a finite number of intervals. The decision tree algorithms first discretize the continuous attributes before proceeding with the attribute selection process. The rule-set learning algorithms exploit discretization methods to produce short and understandable rules. The Bayesian network methods need discrete values to compute conditional probability tables.

Many supervised discretization methods have been proposed in the past. They use a wide variety of criteria based on chi-square [11, 12], entropy [6, 15], impurity measures [5], Minimum Description Length [9]. However, the unsupervised Equal Frequency and Equal Width discretization methods remain attractive methods because of their simplicity. The Equal Width method is used in every statistic program to produce regular histograms. In supervised learning, stacked histograms

are used to visualize the proportion of the class values and their trend related to the descriptive attribute. The Equal Frequency discretization method allows a fair evaluation of the class distribution in each interval. Both Equal Frequency and Equal Width methods have been used to preprocess continuous attributes in benchmarks conducted to evaluate decision tree classifiers or Naïve Bayes classifiers [8, 14, 20]. Dougherty proposes to set the number of bin of Equal Width discretizations to $k = \max(1, 2\log(l))$, where l is the number of distinct values of the attribute. This is a heuristic choice based on examining S-plus histograms. However, in most cases, the number of bins is always set to 10 for the Equal Frequency and Equal Width methods. In the literature, the problem of choosing the optimal number of bins has not been considered in supervised learning. On the opposite, histograms (Equal Width discretizations) have been studied for a long time in the context of unsupervised learning [2, 7, 18, 19]. Histograms are used as non parametric density estimators, and a large number of methods have been suggested to set the optimum number of bins.

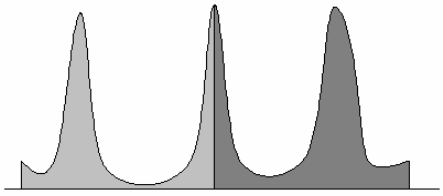

Figure 1: Density function of a continuous attribute with two target classes

To illustrate the main difference between the Equal Width discretization used as a density estimator and a supervised discretization method, we display in figure 1 the density function of a continuous attribute with two target classes. A histogram with a correct bin number properly identifies the three density peaks, but it may mixture the two target classes in the middle peak. On the opposite, a supervised discretization method perfectly separates the target classes by building two intervals, but it is not sensitive to the underlying density. The objective of a good stacked histogram is to consider the target classes in the visualization of the density.

In this paper, we propose a new method to set the optimal number of bins for the Equal Frequency discretization method, when the class information is considered. This method is based on a Bayesian approach. We define a space of discretization models and a prior distribution on this model space. This results in an evaluation criterion of discretizations, which is minimal for the Bayes optimal discretization. We present a search algorithm with super-linear time complexity that allows obtaining optimal discretizations. These results can also be applied to set the optimal number of bins for the Equal Width discretization in supervised learning. We demonstrate through numerous experiments that the optimal Equal Frequency discretization method leads to high quality discretizations.

The remainder of the paper is organized as follows. Section 2 presents the method for finding the optimal bin number in Equal Frequency discretization. Section 3 proceeds with an extensive experimental evaluation.

2 The optimal Equal Frequency discretization method

In this section, we present the approach, the evaluation criterion and the search algorithm used in the optimal method.

2.1 The approach

The objective of the discretization process is to induce a list of intervals that split the numerical domain of a continuous explanatory attribute. The data sample consists of a set of instances described by pairs of values: the continuous explanatory value and the class value. Let n be the number of instances in the data sample and J be the number of classes. If we sort the instances of the data sample according to the continuous values, we obtain a string S of class values.The discretization methods have to solve a problem of model selection, where the data to fit is a string of class values and the model is a discretization model. Let us now recall the principles of the Bayesian approach used as a model selection technique and present its application to select the number of bins in Equal Frequency discretizations.

In the Bayesian approach, the best model is found by maximizing the probability $P(Model/Data)$ of the model given the data. Using the Bayes rule and since the probability $P(Data)$ is constant under varying the model, this is equivalent to maximize:

$$P(Model)P(Data/Model). \quad (1)$$

Once the prior distribution of the models is fixed, the Bayesian approach allows finding the optimal model of the data, provided that the calculation of the probabilities $P(Model)$ and $P(Data/Model)$ is feasible. In definition 1, we introduce a space of discretization models.

Definition 1: A *standard* discretization model is defined by the following properties:

1. the discretization model relies only on the order of the class values in the string S,
2. the discretization model allows to split the string S into a list of substrings (the intervals),
3. in each interval, the distribution of the class values is defined by the frequencies of the class values in this interval.

Such a discretization model is called a SDM model.

Notation:

I: number of intervals
n_i: number of instances in the interval i
n_{ij}: number of instances of class j in the interval i

A SDM model is defined by the set of parameters $\{ I, \{n_i\}_{1\le i\le I}, \{n_{ij}\}_{1\le i\le I,\ 1\le j\le J} \}$.

This definition is very general and most supervised discretization methods rely

on SDM models. They first sort the samples according to the attribute to discretize (property 1) and try to define a list of intervals by partitioning the string of class values (property 2). The evaluation criterion is always based on the frequencies of the class values (property 3). This definition is general enough to include the Equal Frequency discretization models, which are constrained on the way they can split the string into intervals.

2.2 The evaluation criterion

Once a model space is defined, we need to fix a prior distribution on this model space in order to apply the Bayesian approach. We use the universal prior for integers presented in [17] for the choice of the number of intervals in the discretizations. We describe this prior and provide all necessary formulae in appendix. Compared with the uniform prior, the universal prior for integers gives a higher probability to small integers. This characteristic is interesting in the case of discretizations, because discretizations with fewer intervals should be preferred for comprehensibility reasons when several candidate discretizations are equally likely to explain the data. Once the number of interval I is fixed in an Equal Frequency discretization, the bounds of the intervals (parameters n_i) can be derived unconditionally. The remaining n_{ij} parameters to be set are related to the distribution of the class values in the intervals. In the following definition, we propose a prior distribution for the Equal Frequency discretization models.

Definition 2: The following distribution prior on SDM models is called the *Equal Frequency prior*:

1. the number of intervals I is distributed according to the universal prior for integers,
2. for a given number of intervals I, every interval has the same frequency,
3. for a given interval, every distribution of class values in the interval is equiprobable,
4. the distributions of the class values in each interval are independent from each other.

In the last hypothesis of the prior, we have introduced a strong hypothesis of independence of the distribution of the class values. This hypothesis is often assumed (at least implicitly) by many discretization methods, which try to merge similar intervals and separate intervals with significantly different distributions of class values. This is the case for example with the ChiMerge method [12], which merges two adjacent intervals if their distributions of class values are statistically similar (using the chi-square test of independence). For Equal Frequency SDM models, this hypothesis implies that the set of intervals will be searched so that the distributions of class values in each interval are as independent as possible.

Owing to the definition of the model space and its prior distribution, the Bayes formula is applicable to exactly calculate the prior probabilities of the model and the probability of the data given the model. Theorem 1, proven in appendix, introduces a Bayes optimal evaluation criterion.

Theorem 1: A SDM model *M* distributed according to the Equal Frequency prior is Bayes optimal for a given set of instances if the value of the following criterion is minimal:

$$Value(M) = \log_2^*(n) + \sum_{i=1}^{I} \log_2\left(C_{n_i+J-1}^{J-1}\right) + \sum_{i=1}^{I} \log_2\left(n_i! / n_{i,1}! n_{i,2}! .. n_{i,J}!\right). \quad (2)$$

The first term of the criterion corresponds to the choice of the number of intervals, using the universal prior for integers. The second term represents the choice of the class distribution in each interval and the last term encodes the probability of the data given the model. There is no encoding term for the bounds of the intervals since they are unconditionally fixed once the number of intervals is set.

If we change the Equal Frequency prior into an Equal Width prior where every interval has the same width, we obtain exactly the same evaluation criterion.

2.3 The search algorithm

Once the optimality of the evaluation criterion is established, the problem is to design a search algorithm in order to find a discretization model that minimizes the criterion. A straightforward exhaustive search algorithm allows finding the optimal discretization in O(n^2) time. However, the method can be optimized in O(*n.log(n)*) time.

Let us clearly specify the straightforward Equal Frequency discretization method for a given number of intervals *I*. After the instances are sorted according to their descriptive values, the interval boundaries must be chosen so that each interval contains approximately the same number of instances. Let $n_I = \lceil n/I \rceil$ be the mean of the interval frequencies. The leading n_I instances are collected to determine the first interval boundary. If several instances share the same descriptive value, the first interval can contain more than n_I instances. The following intervals are determined with the same method, so that the last interval may contain less than n_I instances. We decide that the last interval will be merged with the preceding interval if its frequency is less than half the mean frequency. All in all, the effective number of intervals can be less than *I*. The time complexity of this algorithm is O(*n.log(n)*) for the sort of the instances.

The calculation of the interval boundaries and the evaluation of the discretization by the Bayes optimal criterion can be done in O(*I*) time. In an initialization step taking O(*n*) time, the cumulated frequencies of the class values can be memorized into an array indexed by the instances. The index of the boundary instances of an interval allow to calculate the class frequencies in the interval owing to the difference of the cumulated frequencies. The evaluation of the discretization criterion thus requires O(*I*) time.

The optimization of the bin number requires the evaluation of all the discretizations for *I* between 1 and *n*. This is the same as the evaluation of all the discretizations for the distinct mean frequencies n_I between 1 and *n*. As $n_I = \lceil n/I \rceil$, we get the following inequality: $I \leq n/(n_I - 1)$ for $n_I > 1$.

The total number of evaluated intervals is then bounded by

$$n + \sum_{n_I=2}^{n} n/(n_I - 1) \sim n(1 + \log(n)).$$

The overall complexity of the optimal algorithm is O(*n.log(n)*) time. This algorithm can also be used to search the optimal Equal Width discretization, provided that the search is restricted to the discretizations having the same numbers of bins that for the Equal Frequency algorithm.

Optimal Equal Frequency algorithm:
- Initialization
  - Sort the explanatory attribute values: O(n.log(n))
  - Compute the cumulated class frequencies for each instance index: O(n)
- Optimization of the discretization
  For I = 1 to n
  - Compute $n_I = \lceil n/I \rceil$; continue if n_I equals n_{I-1}
  - Evaluate the discretization: O(I)
    - For each interval, evaluate the class frequencies as a difference between the cumulated class frequencies
  - Memorize I if better discretization evaluation

3 Experiments

In this section, we present an extensive experimental study both on synthetic and real datasets.

3.1 Noise data experiment

The purpose of this experiment is to evaluate the noise resistance of the discretization method, under varying the sample size. In the case of pure noise data, a robust discretization method should build a single interval, meaning that there is no class information in the explanatory attribute.

The *noise pattern* dataset consists of an explanatory continuous attribute independent from the class attribute. The explanatory attribute is uniformly distributed on the [0, 1] numerical domain and the class attribute consists of two equidistributed class values. The evaluated criterion is the number of unnecessary intervals, since the test accuracy is always 0.5 whatever be the number of intervals in the discretizations. The experiment is done on a large number of sample sizes ranging from 100 instances to 100000 instances. In order to obtain reliable results, it is performed 100000 times for each sample size. Figure 2 presents the average unnecessary interval number obtained by the optimal Equal Frequency discretization method, under varying the sample size. The optimal method is very resistant to noise and almost always produce a single interval. The percentage of multi-interval discretizations is about 10% of the cases for sample size 100. It decreases with the sample size down to about 0.1% for sample size 100000. This behavior seems consistent since the probability of finding an information pattern in

a randomly generated attribute decreases with the sample size.

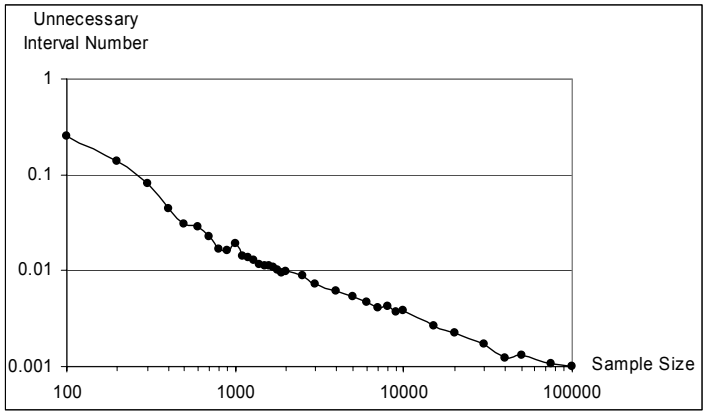

Figure 2: Mean of the unnecessary interval number of the discretizations of an explanatory attribute independent from the class attribute, for the optimal Equal Frequency discretization method

3.2 Real data experiments

3.2.1 Protocol

The purpose of this experiment is to evaluate the predictive quality of the optimal Equal Frequency discretization method on real datasets. In our experimental study, we compare the optimal Equal Frequency and optimal Equal Width methods with the MDLPC method [9] and with the standard Equal Frequency and Equal Width discretization methods using different strategies to set the number of bins.

The MDLPC method is a state of the art supervized discretization method. It exploits a greedy top-down split method, whose evaluation criterion is based on the Minimum Description Length Principle [16]. At each step of the algorithm, the MDLPC method evaluates two hypotheses (to cut or not to cut the interval) and chooses the hypothesis whose total encoding cost (model plus exceptions) is the lowest. The encoding cost is mainly based on the class information entropy of the partition and on the encoding of the position of the cut point. Therefore, the potential diminution of the entropy after a cut is balanced by the encoding cost of the cut point.

The two unsupervised Equal Frequency and Equal Width methods are used with a fixed number of bins set to 10. We also use several strategies proposed to fix the number of bins in histograms (Equal Width discretizations as density estimators), and apply the resulting number of bins both for the Equal Frequency and Equal Width methods. Sturges' rule is a rule of thumbs [19] that consists of taking approximately $1+\log_2(n)$ bins. We can notice that this method is close to the heuristic rule proposed in [8]. Several rules have been proposed to minimize the mean square error between the histogram density estimator and the true density.

These rules generally give a bin width proportional to $n^{-1/3}$. Scott's rule [18] is calibrated with the normal distribution and results in a bin width equal to $3.49\sigma n^{-1/3}$, where σ is the standard deviation of the descriptive attribute. We also use a recent result presented in [2], based on a nonasymptotic evaluation of the performances of penalized maximum likelihood estimators. This method consists of maximizing $L_n(I) - pen(I)$ for $1 \leq I \leq n/\log(n)$, where

$L_n(I) = \sum_{i=1}^{I} n_i \log(n_i I/n)$ is the log-likelihood of the histogram with I bins and

$pen(I) = I - 1 + \log(I)^{2.5}$ is the penalty. This last term is close from Akaike's [1] with a modification derived from [7].

This last method cannot be used to set the number of bins for Equal Frequency discretizations since the criterion is maximum for I=1 when $n_i = n/I$.

To summarize, the evaluated methods are:

- EF(Opt): Equal Frequency with optimal bin number
- EF(Sturges): Equal Frequency with Sturges' bin number
- EF(Scott): Equal Frequency with Scott's bin number
- EF(10): Equal Frequency with 10 bins
- MDLPC: supervized discretization method
- EW(Opt): Equal Width with optimal bin number
- EW(Sturges): Equal Width with Sturges' bin number
- EW(Scott): Equal Width with Scott's bin number
- EW(Birgé): Equal Width with Birgé's bin number
- EW(10): Equal Width with 10 bins

We gathered 15 datasets from U.C. Irvine repository [3], each dataset has at least one continuous attribute and at least a few tens of instances for each class value in order to perform reliable tenfold cross-validations. Table 1 describes the datasets; the last column corresponds to the relative frequency of the majority class. The datasets come from a large variety of knowledge domains, with varying attribute numbers, sample sizes and class numbers. The task in the Adult dataset is to predict if an individual's annual income exceeds $50,000 based on census data. The Australian, Crx and German datasets are dealing with credit approval. The Heart, Hepatitis, Hypothyroid, Diabete and SickEuthyroid originate from the medical domain. The Ionosphere dataset is a problem a radar returns classification. The Iris datasets classifies three classes of iris plants. The Vehicle dataset have to categorize four types of vehicle. The Waveform dataset is a synthetic dataset with noise added to each instance. The Wine dataset purpose is to distinguish three kinds of wine on the basis of their chemical constituents.

Table 1: Datasets

Dataset	Continuous Attributes	Nominal Attributes	Size	Class Values	Majority Class
Adult	7	8	48842	2	76.07
Australian	6	8	690	2	55.51
Breast	10	0	699	2	65.52
Crx	6	9	690	2	55.51
German	24	0	1000	2	70.00
Heart	10	3	270	2	55.56
Hepatitis	6	13	155	2	79.35
Hypothyroid	7	18	3163	2	95.23
Ionosphere	34	0	351	2	64.10
Iris	4	0	150	3	33.33
Diabete	8	0	768	2	65.10
SickEuthyroid	7	18	3163	2	90.74
Vehicle	18	0	846	4	25.77
Waveform	21	0	5000	3	33.92
Wine	13	0	178	3	39.89

3.2.2 Univariate evaluation

In order to evaluate the intrinsic performance of the discretization methods and eliminate the bias of the choice of a specific induction algorithm, we use the protocol presented in [4]. Each discretization method is considered as an elementary inductive method that predicts the local majority class in each learned interval. The discretizations are evaluated for two criteria: accuracy and interval number. The discretizations are performed on the 181 continuous attributes of the datasets, using a stratified tenfold cross-validation. In order to determine whether the performances are significantly different between the optimal Equal Frequency method and the alternative methods, the t-statistics of the difference of the results is computed. Under the null hypothesis, this value has a Student's distribution with 9 degrees of freedom. The confidence level is set to 5% and a two-tailed test is performed to reject the null hypothesis.

The whole result tables are too large to be printed in this paper. The results are summarized in table 2, which reports the mean of the accuracy and interval number per attribute discretization and the number of significant wins and losses.

Table 2: Mean accuracy and interval number for 181 attribute discretizations, number of significant wins and losses for the optimal Equal Frequency method

	Test Accuracy		Interval Number	
	Mean	EF(Opt) wins	Mean	EF(Opt) wins
EF(Opt)	68.4		7.7	
EF(Sturges)	67.4	31/8	8.2	116/46
EF(Scott)	67.0	29/6	13.0	121/40
EF(10)	67.7	29/9	7.3	48/103
MDLPC	68.0	21/14	3.2	1/134
EW(Opt)	67.8	23/12	11.9	100/17
EW(Sturges)	67.2	37/10	9.4	134/35
EW(Scott)	67.2	31/4	14.5	138/31
EW(Birgé)	67.7	24/5	30.0	162/1
EW(10)	67.1	40/9	8.5	124/39

In order to analyze both the accuracy and interval number results, we reported the mean results on a two-criteria plan in figure 3, with the accuracy on the x-coordinate and the interval number on the y-coordinate. We also performed the experiments for the Equal Frequency and Equal Width methods with all bin numbers ranging from 1 to 30 and reported the results on figure 3.

The optimal Equal Frequency method clearly dominates all the other methods. It performs even better than the MDLPC supervised discretization method on the accuracy criterion, at the expense of twice the number of intervals. Both optimal Equal Frequency and Equal Width methods are clearly superior to the matching methods performed with any fixed number of bins. The Equal Width strategies for fixing the number of bins in histograms may perform well as density estimators, but they are not adequate for a good evaluation of the distribution of the class values. The most sophisticated method from [2] works better than the preceding methods from [19] or [18], but it is outperformed by the optimal Equal Width method, both on the accuracy and the number of intervals criteria. The heuristic choices for the Equal Frequency methods derived from the histogram methods are not better than the basic choice of 10 bins.

These results demonstrate that the optimization of the Equal Frequency and Equal Width methods allow to outperform the other strategies used to set the number of bins. The Equal Frequency method should be preferred to the Equal Width method in supervised learning. When the number of bins is optimized, it becomes competitive even with supervised methods.

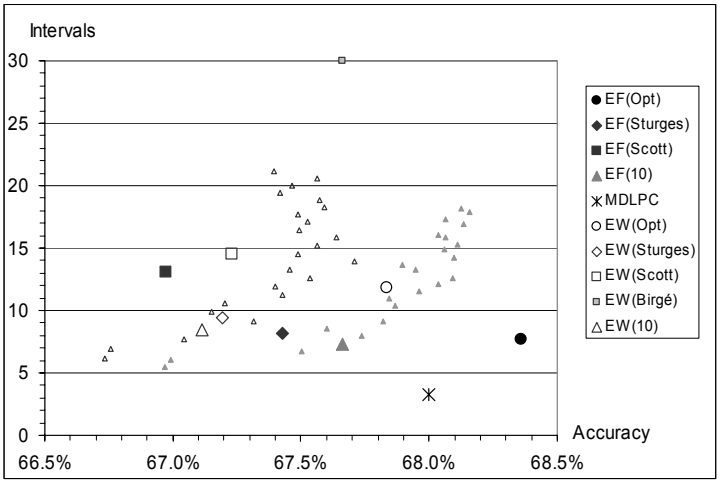

Figure 3: Bi-criteria evaluation of the discretization methods for the accuracy and the interval number

3.2.3 Naïve Bayes evaluation

In order to evaluate the impact of our approach on supervised learning, we conduct a study similar to [8], using the Naïve Bayes classifier. The Naïve Bayes classifier [13] assigns the most probable class value given the explanatory attributes values, assuming independence between the attributes for each class value. The discretization is used as a preprocessing step to estimate the probabilities of the continuous attributes using counts in each interval. In [8], several discretization methods (the MDLPC method, the 1RD method [10] and the Equal Width method with 10 bins) are compared; the evaluation shows that the MDLPC method is the best choice among the tested methods. In our study, we focus on the Equal Frequency and Equal Width methods, and use the MDLPC method for comparison reasons. We set the bin number both to 10 and using our approach.

As the purpose of our experimental study is to compare discretization methods, we chose to ignore the nominal attributes to build the Naïve Bayes classifiers. We ran ten stratified tenfold cross-validations and report the mean and the standard deviation of the test accuracy in table 3. In figure 4, we plot the differences of accuracy between the tested methods and the reference MDLPC method for all datasets.

Table 3: Accuracy of the Naïve Bayes classifier with different discretization methods

	Dataset	EF(Opt)	EW(Opt)	EF(10)	EW(10)	MDLPC
1	Adult	84.50±0.46	84.45±0.44	80.62±0.45	81.39±0.49	84.32±0.45
2	Australian	78.80±4.43	75.72±5.21	79.28±4.11	70.38±4.86	76.94±4.54
3	Breast	97.38±1.94	97.28±1.98	97.44±1.91	97.38±1.93	97.14±2.04
4	Crx	78.57±4.60	76.04±5.00	78.90±4.93	70.32±4.90	77.17±4.91
5	German	72.98±4.12	74.94±3.91	76.01±4.22	75.23±3.82	72.86±3.77
6	Heart	80.41±7.30	80.67±7.78	81.78±7.46	82.00±7.07	80.59±7.72
7	Hepatitis	79.27±10.7	80.34±10.2	80.38±11.2	82.88±10.5	78.68±9.31
8	Hypothyroid	98.21±0.72	98.62±0.59	97.36±0.96	97.46±0.72	98.62±0.52
9	Ionosphere	89.35±4.68	88.89±4.69	89.78±4.60	90.83±4.36	89.58±4.67
10	Iris	93.87±5.63	94.87±5.32	94.13±6.27	95.13±5.15	93.07±5.96
11	Diabete	74.40±4.29	75.27±4.25	75.05±4.18	75.87±4.05	75.51±3.63
12	SickEuthyroid	96.08±1.03	95.50±1.11	93.97±1.23	92.82±1.18	95.97±0.98
13	Vehicle	62.13±4.05	62.55±3.67	60.84±4.46	61.43±4.02	60.81±3.92
14	Waveform	81.02±1.42	80.73±1.52	80.85±1.36	80.86±1.43	80.91±1.50
15	Wine	97.29±3.62	97.35±3.33	97.79±3.58	96.78±4.12	98.48±2.75
	Average	84.28	84.22	84.28	83.38	84.04

The experiment confirms that the MDLPC method obtains better results on average than the Equal Width method with 10 bins. However, this is no longer true when the bin number is optimized by our method. Both Equal frequency discretization methods perform better than the MDLPC methods and the Equal Width methods. The Equal Frequency method with optimized bin number outperforms the MDLPC method on ten datasets. At the 95% confidence level, it is better on five datasets and worse on three. Looking at figure 4, the optimized bin number methods seem to be less prone to the variability of the datasets than the fixed bin number methods.

Overall, the study shows that even simple unsupervised methods can be competitive compared to the state of the art MDLPC method when the target class is considered and when the bin number is optimized.

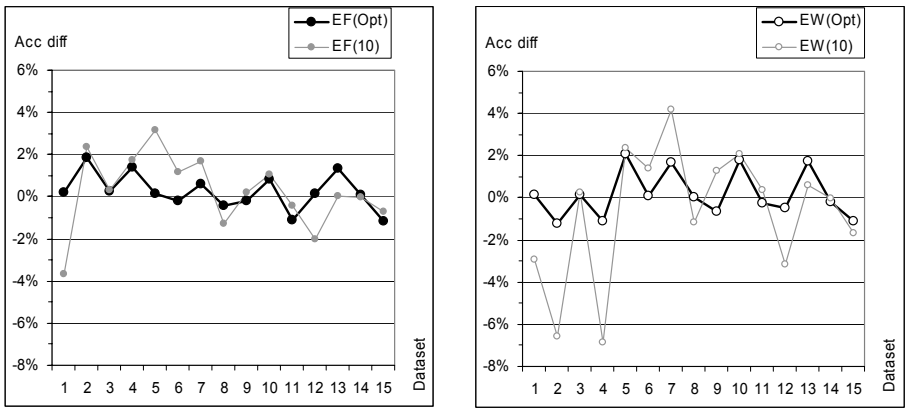

Figure 4: Comparison of the Equal Frequency and Equal Width discretization methods using MDPLC as the reference method, when discretization is used as the preprocessing step for the Naïve Bayes classifier. Graphs indicate the accuracy difference between the four displayed method and the MDLPC method on fifteen datasets

4 Conclusion

The method presented in this paper allows finding the optimal number of bins for the Equal Frequency and Equal Width discretization methods in the context of supervised learning. This method takes advantage of the precise definition of a family of discretization models with a general prior. This provides a new evaluation criterion which is minimal for the Bayes optimal discretization, i.e. the most probable discretization given the data sample. An algorithm is proposed to find the optimal discretization with super-linear time complexity.

Extensive evaluations demonstrate that the method allows building discretizations that are both robust and accurate. The theoretical potential of the method is confirmed on synthetic random data, where the most probable discretization given the data is composed of a single interval. This might be helpful to detect noise attributes and more generally to improve the selection of attributes. The optimal Equal Width method can be used in the exploration step of data mining to produce highly accurate stacked histograms. The optimal Equal Frequency discretization method performs at least as well as the MDLPC supervised method. When used as a preprocessing step of the Naïve Bayes classifier, it obtained better results than the MDLPC method in two third of the datasets used in our evaluation.

When continuous data needs to be discretized, the unsupervised Equal Width and Equal Frequency discretization methods are appealing because of their simplicity. The method presented in this paper keeps these methods simple and makes them become competitive to explore and preprocess continuous data in supervised learning.

Acknowledgments

I am grateful to the anonymous reviewers for their beneficial comments.

References

[1] Akaike, H., A new look at the statistical model identification. IEEE Transactions on Automatic Control 19, 1974, 716-723.
[2] Birgé, L. and Rozenholc, Y., How many bins should be put in a regular histogram, Prépublication 721, Laboratoire de Probabilités et Modèles Aléatoires, Université Paris VI et VII, France, 2002.
[3] Blake, C.L. and Merz, C.J., UCI Repository of machine learning databases [www.ics.uci.edu/~mlearn/MLRepository.html], Irvine, CA: University of California, Department of Information and Computer Science, 1998.
[4] Boullé, M., Khiops: a Discretization Method of Continuous Attributes with Guaranteed Resistance to Noise, In Proceeding of the Third International Conference on Machine Learning and Data Mining in Pattern Recognition, 2003, pp 50-64.
[5] Breiman, J., Friedman, J.H., Olshen, R.A. and Stone, C.J., Classification and Regression Trees, California,: Wadsworth International, 1984.
[6] Catlett, J., On Changing Continuous Attributes into ordered discrete Attributes, In Proceedings of the European Working Session on Learning, Springer-Verlag, 1991, pp 87-102.
[7] Castellan, G., Modified Akaike's criterion for histogram density estimation, Technical Report, Université Paris-Sud, Orsay, 1999.
[8] Dougherty, J., Kohavi, R. and Sahami, M., Supervised and Unsupervised Discretization of Continuous Features, In Proceedings of the 12th International Conference on Machine Learning, San Francisco, CA: Morgan Kaufmann, 1995, pp 194-202.
[9] Fayyad, U. and Irani, K, On the handling of continuous-valued attributes in decision tree generation, Machine Learning 8, 1992, 87-102.
[10] Holte, R.C., Very simple classification rules perform well on most commonly used datasets, Machine Learning 11, 1993, pp 63-90.
[11] Kass, G.V., An explanatory technique for investigating large quantities of categorical data. Applied statistics 29(2), 1980, pp 119-127.
[12] Kerber, R., Chimerge discretization of numeric attributes, In Proceedings of the 10th International Conference on Artificial Intelligence, 1991, pp 123-128.
[13] Langley, P., Iba;, W. and Thompson, K., An analysis of bayesian classifiers. In Proceedings of the 10th National Conference on Artificial Intelligence, AAAI Press, 1992, pp 223-228.
[14] Liu, H., Hussain, F., Tan, C.L. and Dash, M., Discretization: An Enabling Technique, Data Mining and Knowledge Discovery 6(4), 2002, pp 393-423.
[15] Quinlan, J.R., C4.5: Programs for Machine Learning, Morgan Kaufmann, 1993.
[16] Rissanen, J., Modeling by shortest data description, Automatica 14, 1978, pp 465-471.
[17] Rissanen, J., A universal prior for integers and estimation by minimum description length, Ann. Statist. 11, 1983, pp 416-431.
[18] Scott, D.W., Averaged shifted histograms: Effective nonparametric density estimators in several dimensions, Ann. Statist. 13, 1979, pp 1024-1040.
[19] Sturges, H.A., The choice of a class interval, J. Am. Stat. Assoc 21, 1926, pp 65-66.
[20] Yang, Y. and Webb, G.I., A Comparative Study of Discretization Methods for Naïve-Bayes Classifiers, in Proceedings of PKAW 2002, The 2002 Pacific Rim Knowledge

Acquisition Workshop, Tokyo, Japan, 2002, pp 159-173.
[21] Zighed, D.A. and Rakotomalala, R., Graphes d'induction, HERMES Science Publications, 2000, pp 327-359.

Appendix

Universal prior for integers

In this section, we describe and comment the universal prior for integers presented in [13].

When an integer belongs to a finite set of size N, a uniform prior can be used, where each integer has the same probability $1/N$ of appearance. This approach does not apply when the set is infinite, and Rissanen has proposed a universal prior for integers. This universal prior is defined so that the small integers are more probable than the large integers, and the rate of decay is taken to be as small as possible. According to Rissanen, this prior is "universal" because its resulting code length (negative log of the probability) realizes the shortest coding of large integers. This prior is attractive even in the case of finite sets of integers, because it makes small integers preferable to large integers with the slightest possible difference.

The code length of the universal prior for integers is given by

$$L(n) = \log_2(c_0) + \log_2^*(n) = \log_2(c_0) + \sum_{j>1} \max\left(\log_2^{(j)}(n), 0\right), \quad \textbf{(3)}$$

where $\log_2^{(j)}(n)$ is the j[th] composition of $\log_2$ ($\log_2^{(1)}(n) = \log_2(n)$, $\log_2^{(2)}(n) = \log_2(\log_2(n))\ldots$) and $c_0 = \sum_{n>1} 2^{-\log_2^*(n)} = 2.865...$

The universal prior for integers is then $p(n) = 2^{-L(n)}$.

Proof of theorem 1

Theorem 1: A SDM model M distributed according to the Equal Frequency prior is Bayes optimal for a given set of instances if the value of the following criterion is minimal:

$$Value(M) = \log_2^*(n) + \sum_{i=1}^{I} \log_2\left(C_{n_i+J-1}^{J-1}\right) + \sum_{i=1}^{I} \log_2\left(n_i! / n_{i,1}! n_{i,2}! \ldots n_{i,J}!\right). \quad \textbf{(4)}$$

Proof:

The prior probability of a discretization model M can be defined by the prior probability of the parameters of the model $\{ I,\ \{n_i\}_{1\le i\le I},\ \{n_{ij}\}_{1\le i\le I,\ 1\le j\le J}\ \}$.

Let us introduce some notations:

- $p(I)$: prior probability of the interval number I,
- $p(\{n_i\})$: prior probability of the parameters $\{n_1,\ldots,n_I\}$,
- $p(n_i)$: prior probability of the parameter n_i,

- $p(\{n_{ij}\})$: prior probability of the parameters $\{n_{11},...,n_{ij},...,n_{IJ}\}$,
- $p(\{n_{ij}\}_i)$: prior probability of the parameters $\{n_{i1},...,n_{iJ}\}$.

The objective is to find the discretization model M that maximizes the probability $p(M/S)$ for a given string S of class values. Using the Bayes formula and since the probability $p(S)$ is constant under varying the model, this is equivalent to maximize $p(M)p(S/M)$.

Let us first focus on the prior probability $p(M)$ of the model. We have

$$\begin{aligned} p(M) &= p(I, \{n_i\}, \{n_{ij}\}) \\ &= p(I)\, p(\{n_i\}/I)\, p(\{n_{ij}\}/I, \{n_i\}). \end{aligned}$$

The first hypothesis of the equal frequency prior is that the number of intervals is distributed according the universal prior for integers. Thus we get

$$p(n) = 2^{-L(n)} = \frac{1}{c_0} 2^{-\log_2^*(n)}.$$

The second hypothesis is that every admissible partition has the same frequency for a given I. Thus we obtain $p(\{n_i\}/I) = 1$ when the bounds of intervals correspond to I equal frequency intervals and $p(\{n_i\}/I) = 0$ otherwise. In real datasets, it is not always possible to construct intervals having exactly the same frequency. The key point is that for a given number of intervals, the discretization algorithm must provide a unique way of building the intervals.

The last term to evaluate can be rewritten as a product using the hypothesis of independence of the distributions of the class values between the intervals. We have

$$\begin{aligned} p(\{n_{ij}\}/I, \{n_i\}) &= p(\{n_{ij}\}_1, \{n_{ij}\}_2, ..., \{n_{ij}\}_I / I, \{n_i\}) \\ &= \prod_{i=1}^{I} p(\{n_{ij}\}_i / I, \{n_i\}) \\ &= \prod_{i=1}^{I} p(\{n_{ij}\}_i / n_i). \end{aligned}$$

For a given interval i with size n_i, all the distributions of the class values are equiprobable. Computing the probability of one distribution is a combinatorial problem, which solution is:

$$p(\{n_{ij}\}/n_i) = \frac{1}{C_{n_i+J-1}^{J-1}}.$$

Thus,

$$p(\{n_{ij}\}/I, \{n_i\}) = \prod_{i=1}^{I} \frac{1}{C_{n_i+J-1}^{J-1}}.$$

The prior probability of the model is then

$$p(M) = 2^{-L(n)} \prod_{i=1}^{I} \frac{1}{C_{n_i+J-1}^{J-1}}.$$

Let us now evaluate the probability of getting the string S for a given model M.

We first split the string S into I sub-strings S_i of size n_i and use again the independence assumption between the intervals. We obtain

$$\begin{aligned} p(S/M) &= p\left(S/I,\{n_i\},\{n_{ij}\}\right) \\ &= p\left(S_1,S_2,...,S_I/I,\{n_i\},\{n_{ij}\}\right) \\ &= \prod_{i=1}^{I} p\left(S_i/I,\{n_i\},\{n_{ij}\}\right) \\ &= \prod_{i=1}^{I} \frac{1}{\left(n_i!/n_{i,1}!n_{i,2}!...n_{i,J}!\right)}, \end{aligned}$$

as evaluating the probability of a sub-string S_i under uniform prior turns out to be a multinomial problem.

Taking the negative log of the probabilities, the maximization problem turns into the minimization of the claimed criterion

$$Value(M) = \log_2^*(n) + \sum_{i=1}^{I} \log_2\left(C_{n_i+J-1}^{J-1}\right) + \sum_{i=1}^{I} \log_2\left(n_i!/n_{i,1}!n_{i,2}!...n_{i,J}!\right).$$