Comparative Analysis of Continuous Entropy Estimation with Different Unsupervised Discretization Methods

Jian Fang, Li-Na Sui, Hong-Yi Jian
Mathematics and Computer Department, Hebei Normal University for Nationalities, Chengde 067000, Hebei, China
Email: csjfang@yeah.net, cslnsui@yeah.net

Abstract—**In this paper, we compare and analyze the performances of nine unsupervised discretization methods, i.e., equal width discretization (EWD), equal frequency discretization (EFD), k-means clustering discretization (KMCD), ordinal discretization (OD), fixed frequency discretization (FFD), non-disjoint discretization (NDD), proportional discretization (PD), weight proportional discretization (WPD), mean value and standard deviation discretization (MVSDD), based on the application of continues entropy estimation. Firstly, we give the detailed description about the concept of continuous entropy estimation. Then, we introduce the nine different unsupervised discretization methods. Finally, we conduct the estimation of continuous entropy based on 15 probability density distributions, i.e., Beta, Cauchy, Central Chi-Squared, Exponential, F, Gamma, Laplace, Logistic, Lognormal, Normal, Rayleigh, Student's-t, Triangular, Uniform, and Weibull distributions. The experimental results show that in comparison with the sophisticated discretization methods-OD, FFD, NDD, PD, and WPD, EWD and EFD can also the considerable estimation performances. Moreover, we also illustrate the relationship between the size of training dataset and the estimation performance.**

Keywords—continuous entropy estimation; probability density distribution; unsupervised discretization

I. Introduction

Entropy which has been proposed by C.E. Shannon [1], [2] is always used to measure the amount of information contained in a certain domian. It has a wide application in many fields, e.g., decision tree generation [3], [4] and feature subsect selection [5], [6], etc. For a specific learning problem, the dataset is composed with several features (i.e., variables) of which the values are discrete or continuous. "Discrete" refers to the variables taking on categorical values and "continuous" refers to the variables taking on integer or real values.

In statistics and machine learning, discretization refers to the process of converting or partitioning continuous attributes, features or variables to discretized or nominal attributes, features, or variables. Over the years, many discretization methods have been proposed and tested to show that discretization helps improve the performance of learning methods and helps understand the learning result. One of taxonomies [7] is to classify primary discretization methods into supervised discretization and unsupervised discretization, where supervised discretization uses the class or label information to select the discretization cut points and unsupervised discretization determines the cut points without the usage of class or label information. In the setting of entropy estimation without class information provided, supervised discretization may not be competent to implement the entropy computation for continuous variables. So, unsupervised discretization is considered as a capable candidate to estimate the continuous entropy.

In our study, nine common unsupervised discretization methods are introduced and employed as the competitors. Equal width discretization (EWD) [8] and equal frequency discretization (EFD) [9] are two mostly used and simplest methods. The experimental observations in numerous literatures show that the satisfactory performances and reasonable effectiveness of EWD and EFD are not affected by their directness and simplicity. K-means clustering discretization (KMCD) [10] uses k-means clustering [11] to determine intervals for the discrete variables. Ordinal discretization (OD) [12], [13] aims at taking advantage of the ordering information implicit in the continuous variables, so the ordering information of continuous variables is preserved when a transformation of discretized data is carried out. Fixed frequency discretization (FFD) [7], non-disjoint discretization (NDD) [14], proportional discretization (PD) [15], and weight proportional discretization (WPD) [16] are designed intentionally for managing the bias and variance generated during the discretization of continuous variables. The gratifying experimental results have been reported when these four discretization methods are applied to naive Bayesian classifier [17], [18]. Mean value and standard deviation discretization (MVSDD) [6] are applied to feature selection and the better experimental results are obtained when continuous variables are discretized by MVSDD.

In this paper, we compare and analyze the performances of these nine unsupervised discretization methods mentioned above based on the application of continues entropy estimation. Firstly, we give the detailed description about the concept of continuous entropy estimation. Then, we introduce the nine different unsupervised discretization methods. Finally, we conduct the estimation of continuous entropy based on 15 probability density distributions, i.e., Beta, Cauchy, Central Chi-Squared, Exponential, F, Gamma, Laplace, Logistic, Lognormal, Normal, Rayleigh, Student's-t, Triangular, Uniform, and Weibull distributions. The experimental results show that in comparison with the sophisticated discretization methods-

Published by Atlantis Press, Paris, France.

TABLE I
THE DETAILED INFORMATION OF THESE NINE DISCRETIZATION METHODS

Discretization method	The number of intervals	The width of interval	Note
EWD	k	$w = \frac{x_{\max} - x_{\min}}{k}$	$x_{\min} = \min\{x_1, x_2, \cdots, x_n\}$, $x_{\max} = \max\{x_1, x_2, \cdots, x_n\}$, and k is a user predefined parameter.
EFD	k	$\lceil n/k \rceil$	$\lceil u \rceil$ denotes the rounding of the element u to the nearest integers towards infinity.
KMCD	k	$\lceil n/k \rceil$	The parameter k is determined by using the k-means clustering technology.
OD	-	-	OD firstly discretizes by using some primary discretization method (e.g., EWD, EFD, or KMCD). Then, the discretized attribute $\mathrm{X}^* = \{x_1^*, x_1^*, \cdots, x_k^*\}$ is split with $\mathrm{X}_j^*, (j = 1, 2, \cdots, k-1)$.
FFD	$\lceil n/m \rceil$	30	The empirical results show when $m = 30$, the better performance can be obtained in naive Bayesian classifier context by managing the discretization bias and variance.
NDD	$k' = \frac{n}{m'}$	$m' = \frac{m}{3}$	Firstly, the k' "atomic intervals" need to be generated and every "atomic interval" contains m' continuous observations. Then, a total of k "actual intervals" can be constructed by combining three consecutive atomic intervals.
PD	$\sqrt{n}$	$\sqrt{n}$	PD aims to resolve the conflict between variance and bias by setting the interval size and number which are proportional to the number of training instances.
WPD	$\frac{n}{m_{\min}}$	$\frac{m_{\min} + \sqrt{m_{\min}^2 + 4n}}{2}$	WPD weighs discretization variance reduction more than bias reduction by setting a minimum interval size $m_{\min} = 30$ to make the probability estimation more reliable.
MVSDD	3	-	The cut points of discretized intervals are $\mu - \alpha\sigma$ and $\mu + \alpha\sigma$.

OD, FFD, NDD, PD, and WPD, EWD and EFD can also the considerable estimation performances. Moreover, we also illustrate the relationship between the size of training dataset and the estimation performance.

II. The Continuous Entropy Estimation

It is well acknowledged that Shannon entropy [1], [2] can be implemented sophisticatedly and efficiently for the discrete variables as the following Eq. (1):

$$\mathrm{H}(\mathrm{X}) = -\sum_{\mathrm{i}=1}^{n} p(x_i) \operatorname{In}[p(x_i)], \tag{1}$$

where, let X be a discrete random variable taking a finite number of possible values x_1, x_2, $\cdots$, x_n with probabilities $p(x_1)$, $p(x_2)$, $\cdots$, $p(x_n)$ respectively such that $p(x_i) \geq 0$, $i = 1, 2, \cdots, n$ and $\sum_{i=1}^{n} p(x_i) = 1$. $\operatorname{In}(u)$, $u > 0$ is the natural logarithm which is the logarithm to the base e, where e is an irrational constant approximately equal to 2.718. Note that $p\operatorname{In}(p) = 0$ when $p = 0$. However, many learning tasks are often involved with the continuous variables. The mathematical formula for continuous variables can be summarized as following Eq. (2) by extending the discrete entropy to continuous case:

$$\mathrm{H}(\mathrm{X}) = -\int_{-\infty}^{+\infty} f(x) \operatorname{In}[f(x)]\, dx, \tag{2}$$

where, let X be a continuous random variable taking the probability density function $f(x)$ such that $\int_{-\infty}^{+\infty} f(x)\, dx = 1$. From the Eq. (2), we can find that there are two main handicaps when the entropy computation for continuous variables is implemented: the unknown of probability density function and the evaluation of integral paradigm. The main strategy for overcoming these two difficulties in the entropy computation of continuous variables is to discretize the continuous variables into discrete ones and then calculate the discrete entropy according to the Eq. (1).

III. Nine Unsupervised Discretization Methods

In [19], He et al. summarized the nine different unsupervised discretization methods. Different from their work, we introduce these nine methods from the viewpoint of characters of discretized intervals, including the number, width, specification of intervals. We summarize these characters in TABLE I. For OD in TABLE I, we use the following Eq. (3) to estimation the continuous entropy:

$$\mathrm{H}(\mathrm{X}) \approx \mathrm{H}(\mathrm{X}^*) = \frac{\sum_{j=1}^{k-1} \mathrm{H}(\mathrm{X}_j^*)}{k-1}. \tag{3}$$

Meanwhile, we also consider the computational complexities of these nine discretization methods. the computational complexities of EWD, EFD, OD, FFD, NND, PD, WPD, and MVSDD are $O(n \log_2 n)$. And, the complexity of KMCD is $O\left(n^{k+1} \log_2 n\right)$, where k is number of clusters.

IV. Experimental Setup and Analysis

In this paper, in order to compare the estimation performances of above-mentioned nine discretization methods, we

TABLE II
THE BENCHMARK 15 PROBABILITY DENSITY DISTRIBUTIONS

Num	Distribution	Density function	Continuous entropy value	Support interval
1	Beta	$f\left(x\right)=\frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha,\beta)},\alpha\succ 0,\beta\succ 0$	$\mathrm{In}\left[B\left(\alpha,\beta\right)\right]-\left(\alpha-1\right)\Psi\left(\alpha\right)-\left(\beta-1\right)\Psi\left(\beta\right)$ $+\left(\alpha+\beta-2\right)\Psi\left(\alpha+\beta\right)$	$x\in[0,1]$
2	Cauchy	$f\left(x\right)=\frac{1}{\pi}\left(\frac{\lambda}{x^2+\lambda^2}\right),\lambda\succ 0$	$\mathrm{In}\left(4\pi\lambda\right)$	$x\in(-\infty,+\infty)$
3	Chi-Squared	$f\left(x\right)=\frac{2^{-\frac{k}{2}}x^{\frac{k}{2}-1}e^{-\frac{x}{2}}}{\Gamma(k/2)},k\succ 0$	$\mathrm{In}\left[2\Gamma\left(k/2\right)\right]+\left(1-\frac{k}{2}\right)\Psi\left(k/2\right)+\frac{k}{2}$	$x\in[0,+\infty)$
4	Exponential	$f\left(x\right)=\lambda\exp\left(-\lambda x\right),\lambda\succ 0$	$1-\mathrm{In}\left(\lambda\right)$	$x\in[0,+\infty)$
5	F	$f\left(x\right)=\frac{n_1^{\frac{n_1}{2}}n_2^{\frac{n_2}{2}}}{B\left(\frac{n_1}{2},\frac{n_2}{2}\right)}\frac{x^{\frac{n_1}{2}-1}}{(n_2+n_1x)^{\frac{n_1+n_2}{2}}},$ $n_1\succ 0,n_2\succ 0$	$\mathrm{In}\left[\frac{n_1}{n_2}B\left(\frac{n_1}{2},\frac{n_2}{2}\right)\right]+\left(1-\frac{n_1}{2}\right)\Psi\left(\frac{n_1}{2}\right)-\left(1+\frac{n_2}{2}\right)\Psi\left(\frac{n_2}{2}\right)$ $+\frac{n_1+n_2}{2}\Psi\left(\frac{n_1+n_2}{2}\right)$	$x\in[0,+\infty)$
6	Gamma	$f\left(x\right)=x^{k-1}\frac{e^{-x/\theta}}{\theta^k\Gamma(k)},k\succ 0,\theta\succ 0$	$k+\mathrm{In}\theta+\mathrm{In}\left[\Gamma\left(k\right)\right]+\left(1-k\right)\Psi\left(k\right)$	$x\in[0,+\infty)$
7	Laplace	$f\left(x\right)=\frac{1}{2b}\exp\left(-\frac{\lvert x-\mu\rvert}{b}\right),\mu\in R,b\succ 0$	$\mathrm{In}\left(2b\right)+1$	$x\in(-\infty,+\infty)$
8	Logistic	$f\left(x\right)=\frac{1}{4s}\mathrm{sech}^2\left(\frac{x-\mu}{2s}\right),\mu\in R,s\succ 0$	$\mathrm{In}\left(s\right)+2$	$x\in(-\infty,+\infty)$
9	Lognormal	$f\left(x\right)=\frac{1}{x\sigma\sqrt{2\pi}}\exp\left[-\frac{(\mathrm{In}x-\mu)^2}{2\sigma^2}\right],$ $\mu\in R,\sigma^2\succ 0$	$\frac{1}{2}+\frac{1}{2}\mathrm{In}\left(2\pi\sigma^2\right)+\mu$	$x\in(0,+\infty)$
10	Normal	$f\left(x\right)=\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right],$ $\mu\in R,\sigma^2\succ 0$	$\mathrm{In}\left(\sigma\sqrt{2\pi e}\right)$	$x\in(-\infty,+\infty)$
11	Rayleigh	$f\left(x\right)=\frac{x}{\sigma^2}\exp\left(-\frac{x^2}{2\sigma^2}\right),\sigma\succ 0$	$1+\mathrm{In}\left(\frac{\sigma}{\sqrt{2}}\right)+\frac{\gamma_E}{2}$	$x\in[0,+\infty)$
12	Student's-t	$f\left(x\right)=\frac{\left(1+x^2/v\right)^{-\frac{v+1}{2}}}{\sqrt{v}B(1/2,v/2)},v\succ 0$	$\frac{v+1}{2}\left[\Psi\left(\frac{v+1}{2}\right)-\Psi\left(\frac{v}{2}\right)\right]+\mathrm{In}\left[\sqrt{v}B\left(\frac{1}{2},\frac{v}{2}\right)\right]$	$x\in(-\infty,+\infty)$
13	Triangular	$f\left(x\right)=\begin{cases}\frac{2(x-a)}{(b-a)(c-a)}, & a\leq x\leq c\\ \frac{2(b-x)}{(b-a)(b-c)}, & c\leq x\leq b\end{cases}$	$\mathrm{In}\left(\frac{b-a}{2}\right)+\frac{1}{2}$	$x\in[a,b]$
14	Uniform	$f\left(x\right)=\frac{1}{b-a}$	$\mathrm{In}\left(b-a\right)$	$x\in[a,b]$
15	Weibull	$f\left(x\right)=\frac{k}{\lambda}\left(\frac{x}{\lambda}\right)^{k-1}e^{-(x/\lambda)^k},$ $\lambda\succ 0,k\succ 0$	$\gamma_E\left(1-\frac{1}{k}\right)+\mathrm{In}\left(\frac{\lambda}{k}\right)+1$	$x\in[0,+\infty)$

select 15 true probability density distributions [19], [20] as our testing-bed. The detailed information (including he density functions, the continuous entropy values, and the corresponding support intervals) of these 15 probability distributions are listed in TABLE II.

Our experimental comparison is arranged as following procedures. Firstly, for every distribution with different dataset size 10, 20, 30, $\cdots$, 250, we randomly generate 250 samples. Secondly, the estimated continuous entropy is computed according to Eq. (1) by using these nine different discretization methods. Thirdly, we record the estimated error. Fourthly, the above-mentioned three procedures are repeated 100 times and the average errors are plotted in Fig. 1.

Form Fig. 1, we can find that in comparison with the sophisticated discretization methods-OD, FFD, NDD, PD, and WPD, EWD and EFD can also the considerable estimation performances. Because with the increasing of training samples, the estimation performances of OD, FFD, NDD, PD, and WPD all keep the trends of gradual increase after the training sample reach some specific size. That is to say, OD, FFD, NDD, PD, and WPD are inappropriate to compute the continuous entropy for dataset with the large size.

V. Conclusion

In this paper, we compare and analyze the performances of nine unsupervised discretization methods, i.e., EWD, EFD, KMCD, OD, FFD, NDD, PD, WPD, and MVSDD, in the application of continues entropy estimation based on 15 probability density distributions, i.e., Beta, Cauchy, Central Chi-Squared, Exponential, F, Gamma, Laplace, Logistic, Lognormal, Normal, Rayleigh, Student's-t, Triangular, Uniform, and Weibull distributions. The experimental results give our an important and useful insight to the application of these discretization methods when computing the continuous entropy.

References

[1] C. E. Shannon, "A mathematical theory of communication," The Bell System Technical Journal, vol. 27, pp. 379-423, 623-656, 1948.

[2] C. E. Shannon, "Prediction and entropy of printed English," The Bell System Technical Journal, vol. 30, pp. 50-64, 1951.

[3] X. Z. Wang, C. R. Dong, "Improving generalization of fuzzy if-then rules by maximizing fuzzy entropy," IEEE Transactions on Fuzzy Systems, vol. 17, no. 3, pp. 556-567, 2009.

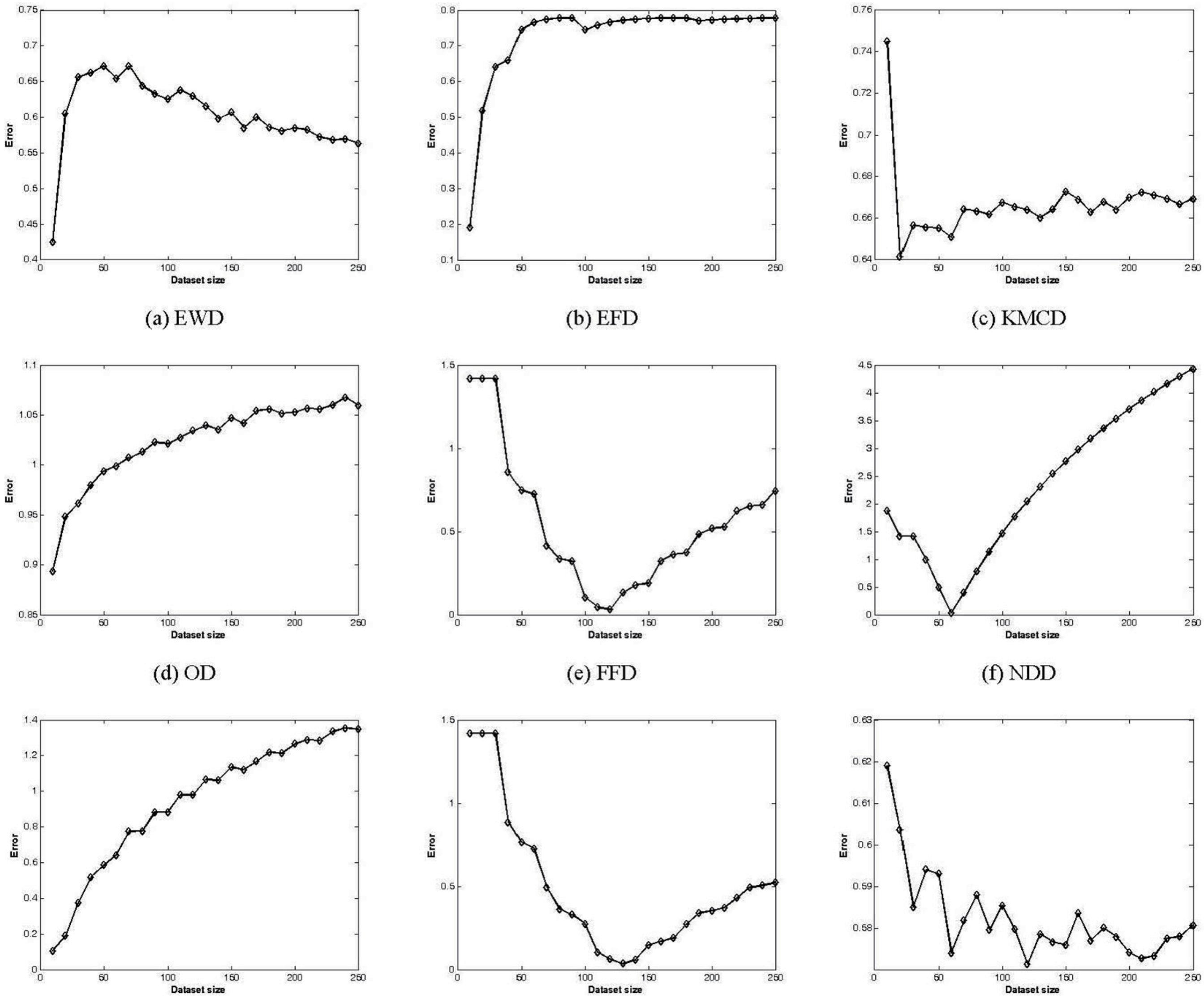
[4] W. G. Yi, M. Y. Lu, Z. Liu, "Multi-valued attribute and multi-labeled data decision tree algorithm," International Journal of Machine Learning and Cybernetics, vol. 2, no. 2, pp. 67-74, 2011.

[5] F. Fleuret, "Fast Binary Feature selection with conditional mutual information," Journal of Machine Learning Research, vol. 5, 1531-1555, 2004.

[6] H. C. Peng, F. H. Long, C. Ding, "Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 27, no. 8, pp. 1226-1238, Aug 2005.

[7] Y. Yang, G. I. Webb, "Discretization for Naive-Bayes Learning Managing Discretization Bias and Variance," Machine Learning, vol. 74, no. 1, pp. 39-74, 2009.

[8] J. Catlett, "On Changing Continuous Attributes into Ordered Discrete Attributes," Lecture Notes in Computer Science, vol. 482, pp. 164-178, 1991.



(a) EWD

(b) EFD

(c) KMCD

(d) OD

(e) FFD

(f) NDD

(g) PD

(h) WPD

(i) MVSDD

Figure 1. The performances of different discretization methods on continuous entropy estimation

[9] R. Kerber, “ChiMerge: Discretization of Numeric Attributes,” In Proceedings of the Tenth National Conference on Artificial Intelligence, pp.123-128, 1992.

[10] L. Torgo, J. Gama, “Search-Based Class Discretization,” Lecture Notes in Computer Science, vol. 1224, pp. 266-273, 1997.

[11] J. A. Hartigan, M. A. Wong, “Algorithm AS 136: A K-Means Clustering Algorithm,” Journal of the Royal Statistical Society-Series C: Applied Statistics, vol. 28, no. 1, pp. 100-108, 1979.

[12] E. Frank, I. H. Witten, “Making Better Use of Global Discretization,” In Proceedings of the 16th International Conference on Machine Learning, pp. 115-123, 1999.

[13] S. A. Macskassy, H. Hirsh, A. Banerjee, A. A. Dayanik, “Using Text Classifiers for Numerical Classification,” In Proceedings of the Seventeenth International Joint Conference on Artificial Intelligence, pp. 885-890, 2001.

[14] Y. Yang, G. I. Webb, “Non-Disjoint Discretization for Naive-Bayes Classifiers,” In Proceedings of the 19th International Conference on Machine Learning, pp. 666-673, 2002.

[15] Y. Yang, G. I. Webb, “Proportional k-Interval Discretization for Naive-Bayes Classifiers,” In Proceedings of the 12th European Conference on Machine Learning, pp. 564-575, 2001.

[16] Y. Yang, G. I. Webb, “Weighted Proportional k-Interval Discretization for Naive-Bayes Classifiers,” In Proceedings of the Seventh Pacific-Asia Conference on Knowledge Discovery and Data Mining, pp. 501-512, 2003.

[17] Y. Yang, G. I. Webb, “A Comparative Study of Discretization Methods for Naive-Bayes Classifiers,” In Proceedings of the 2002 Pacific Rim Knowledge Acquisition Workshop in PRICAI 2002, pp. 159-173, 2002.

[18] Y. Yang, “Discretization for Naive-Bayes Learning,” The School of Computer Science and Software Engineering of Monash University, 2003.

[19] Y. L. He, J. N. K. Liu, X. Z. Wang, Y. X. Hu, “Optimal bandwidth selection for re-substitution entropy estimation,” Applied Mathematics and Computation, vol. 219, no. 8, pp. 3425-3460, 2012.

[20] A. V. Lazo, P. Rathie, “On the Entropy of Continuous Probability Distributions,” IEEE Transactions on Information Theory, vol. 24, no. 1, pp. 120-122, 1978.