Available online www.ejaet.com

European Journal of Advances in Engineering and Technology, 2017, 4 (6): 488-491

Research Article ISSN: 2394 - 658X

A Comparative Analysis of Some Discretization Techniques Using Mutual Information and Transfer Entropy

Barenya Bikash Hazarika and Syed Sazzad Ahmed

Department of Commuter Science Engineering & Information Technology,
Assam Don Bosco University, Guwahati, India
barenya1431@gmail.com

ABSTRACT

In data-mining, many algorithms can't process continuous value, so, discretization is an important method in data-mining application. Discretization describes converting some continuous data into some discrete values. In our paper firstly, we will try to implement three discretization techniques on a dataset, namely, Equal Width, Global Equal Width and Equal Frequency. Secondly we will apply two causality finding techniques namely, Mutual Information and Transfer Entropy and compare the performance among those discretization techniques.

Keywords: Discretization, Causality, Performance Evaluation Techniques, Mutual Information, Transfer Entropy

INTRODUCTION

Data may be of various types: Continuous or discrete or nominal. Discretization concerns with the method of converting non-discrete functions, models, and equations into discrete values. Without discretization, the learning will be less efficient. This process is also called as quantization. Simply, discretization means converting some continuous data to discretized data. There are several techniques for discretization [1]. Three among them are-

Equal Width Discretization

Equal-width interval discretization is the simplest technique of discretization that divides the continuous values into some equal sized bins, k, where k =1, 2, 3, etc. The process finding the minimum, Vmin and maximum, Vmax, values and a threshold value. The values less than the threshold are replaced as 0 and the values greater than or equal to the threshold values are replaced as 1 [2-3]. In simple words, this technique, first, confirms the maximum and minimum of the numeric data, and divides it into some equal-width intervals which are discrete [4]. In our work, we took the size of bins as 2.

Global Equal Width Discretization

The basic difference between Equal Width Discretization and Global Equal Width discretization technique is that, here, global data is considered for calculation, but, in equal width discretization technique, local data is considered. For example, if we take a column of a dataset for operation then it will be called as equal width discretization and if the whole dataset is considered then it is called as global equal width discretization technique i.e. Equal Width Discretization is for local region and Global Equal Width is for global region.

Equal Frequency Discretization

The equal-frequency discretization technique determines the lowest (min) and highest (max) values of the discretized values by converting them into some categorical variables. Firstly, it sorts the values in ascending order. After that, it separates the values into k intervals (k is user given) and makes sure that every interval is consist of the same amount of sorted values. For equal frequency, if a continuous value happens many time, it could result as the happenings to be allocated into distinct bins. This method tries to succeed in dealing with the limitations of the equal-width interval discretization by dividing the continuous attribute in same number of instances. This technique is also called as k-interval technique of discretization [5-6]. Simply, the Equal Frequency method confirms the minimum and maximum of the given numeric values, and separate the range into some intervals (k) which consist of the same number of sorted values in ascending order.

CAUSALITY AND CAUSALITY FINDING TECHNIQUES

Causality depicts the relation between the cause and its effect. Though cause and effect are deterministic in nature, but, it involves probability language [7]. Causality finding techniques are of various types. Few of the techniques that we are going to use are-

Mutual Information (MI)

In probability theory and information theory, MI defines as the normal measure of dependence between a couple variables. If x and y are two random variables, one of the most basic questions that arise is the mutual information between them. It is the 'amount of information' which is acquired about a random variable x, from the other random variable y. The main idea of mutual information is directly linked to that of entropy of a random variable [8-10]. The common measure of MI is bit.

If x and y are two random variables, then, Mutual information is the amount of information that transfers from x to y [11]. The main disadvantage of mutual information is that it has no directions because directed information tells more things about the structure.

Transfer Entropy (TE)

Transfer entropy, was developed within the Physical Sciences community as a means of testing for directional influence in complex systems. The development of the concept of Transfer Entropy was done by Schreiber, who quantified the amount of information from one-time series data to another. Transfer entropy from a process x to other process y is the amount of unknown values are lessened in future values of y by knowing the past values of x given past values of y. TE is a MI with the history of the effected variable in the condition.

The TE measures the amount of information transferred from one variable *x* to another variable *y* and also variable y to variable x [12-14]. If x and y are two random processes then, $T_{x\text{->}y}$ describes the transfer entropy from x to y and $T_{y\text{->}x}$ describes the transfer entropy of y to x. The common measurement unit of TE is bit [15].

PERFORMANCE EVALUATION METHODS

Precision

Precision which is also called as Confidence in Data Mining depicts that how much of the Predicted Positive cases are correctly Real Positives. If the precision is high, it means that the algorithm gives more relevant results [16-17].

Sensitivity

Recall which is called as Sensitivity in Psychology is the portion of Real Positive cases that are correctly predicted as positive. High recall suggest that the algorithm, we are using is giving us the most of the relevant results [16-17].

$F_\beta$score

It is the measure of test accuracy, considers both precision and recall. It can also be called as Harmonic Mean of Precision and Recall. [18-19]

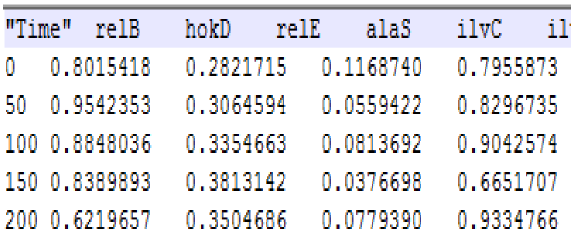
If ß > 1, F becomes more recall-oriented and if ß < 1, it becomes more precision-oriented. We took ß= 0.5 in our calculations. It reaches its optimal value at 1 and worst value at 0 [20].

ACC

Accuracy describes the overall effectiveness.

DATASET

The dataset we have considered here is obtained from DREAM (Dialogue on Reverse Engineering Assessment & Methods) challenge [21]. In the dataset, the X axis (horizontally) describes the concentration of genes at various time. From the figure, relB, hokD, relE etc. are different types of genes. Y axis (vertically) describes the time starting from 0 millisecond. The name of the dataset is Ecoli 2.



"Time"	relB	hokD	relE	alaS	ilvC	il
0	0.8015418	0.2821715	0.1168740	0.7955873		
50	0.9542353	0.3064594	0.0559422	0.8296735		
100	0.8848036	0.3354663	0.0813692	0.9042574		
150	0.8389893	0.3813142	0.0376698	0.6651707		
200	0.6219657	0.3504686	0.0779390	0.9334766		

Fig.1 Portion of the Ecoli-2 dataset [21]

STEPS OF IMPLEMENTATION

Step1: Firstly, we will take a continuous dataset (Time- Series Data) as an input
Step2: Secondly, we will discretize the continuous data using three discretization techniques
Step3: After that we will implement Mutual Information and transfer entropy on that discretized data for each method to get the inferred network.
Step4: We will compare the inferred network using various performance evaluation techniques with the gold standard dataset like Fscore, ACC and AUC.

RESULTS AND ANALYSIS

Table -1 Calculation of tp, fp, tn and fn

	Discretization	fp	fn	tp	tn
Mutual Information	EW	2	14	0	209
	GEW	9	13	1	202
	EF	29	13	1	182
Transfer Entropy	EW	68	9	5	143
	GEW	53	11	3	158
	EF	64	6	8	147

Table -2 Results Obtained by using Performance Evaluation Methods

	Discretization	Sensitivity	Specificity	Precision	F_βScore	Accuracy(ACC)	AUC
MI	EW	0	0.9905	0	0	0.9288	0.4952
	GEW	0.0714	0.9573	0.1	0.9259	0.9022	0.5143
	EF	0.0714	0.0865	0.0333	0.0373	0.8133	0.4669
TE	EW	0.3571	0.6777	0.0684	0.0816	0.0630	0.1360
	GEW	0.2142	0.7488	0.0535	0.6577	0.7155	0.6888
	EF	0.5714	0.6966	0.1111	0.5174	0.4815	0.6340

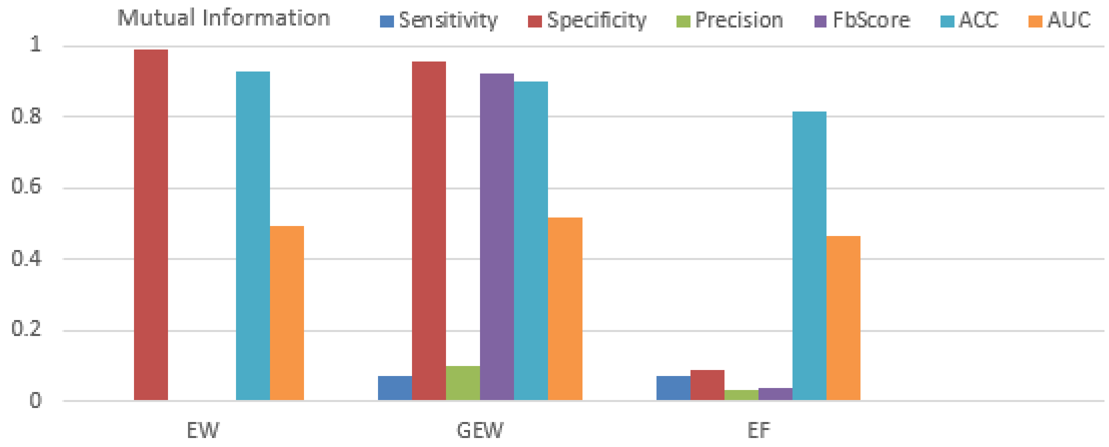


Fig.2 The obtained results after applying Mutual Information

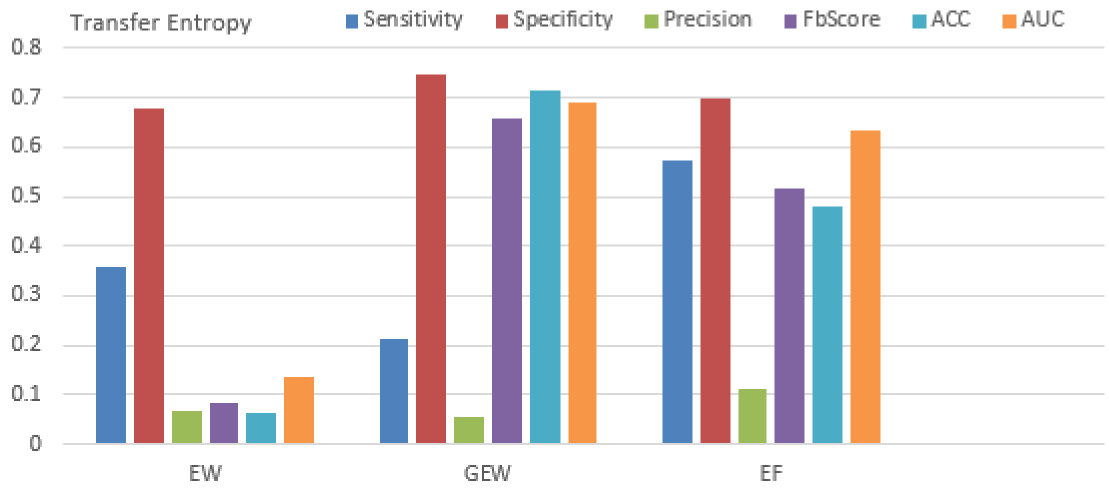


Fig.3 The obtained results after applying Transfer Entropy

We have found that –

i) For Sensitivity	a) MI GEW=EF>EW b) TE EF>EW>GEW	ii) For Specificity	a) MI EW>GEW>EF b) TE GEW>EF>EW	iii) Precision	a) MI GEW>EF>EW b) TE EF>EW>GEW
iv) For F_{β}Score	a) MI GEW>EF>EW b) TE GEW>EF>EW	v) For ACC	a) MI EW>GEW>EF b) TE GEW>EF>EW	vi) For AUC	a) MI GEW>EW>EF b) TE GEW>EF>EW

CONCLUSION

In our work, we have discretized the continuous or time series datasets using three discretization techniques namely equal width, global equal width and equal frequency discretization technique and we have applied two causality finding techniques- MI and TE on the discretized datasets to get the inferred network. We compared the inferred network with the gold standard network using some performance evaluation techniques like Sensitivity, Specificity, Precision, F_{β} Score, ACC and AUC. For MI we got the best performance in Equal Width discretization taking accuracy as the performance measure. For TE we got the best performance in Global Equal Width discretization taking accuracy as the performance measure. Among the two causality finding techniques TE outperformed MI whereas Global Equal Width gave better result compared to Equal Width and Equal Frequency discretization.

REFERENCES

[1] R Dash, RL Paramguru and R Dash, Comparative Analysis of Supervised and Unsupervised Discretization Techniques, International Journal of Advances in Science and Technology, **2011**, 2, 29-37.

[2] R Baeza-Yates and B Riberio-Neto, Modern Information Retrieval, Addison *Wesley*, **1999**,463.

[3] K Coussement, S Lessmann and G Verstraeten, A Comparative Analysis of Data Preparation Algorithms for Customer Churn Prediction: A Case Study in the Telecommunication Industry, *Decision Support Systems*, **2017**, 95, 27-36.

[4] X He, F Min and W Zhu, A Comparative Study of Discretization Approaches for Granular Association Rule Mining, *IEEE Canadian Conference of Electrical and Computer Engineering*, Regina, Canada, **2013**, 1-5.

[5] F Muhlenbach and R Rakotomalala, Discretization of Continuous Attributes, *Encyclopedia of Data Warehousing and Mining*, **2005**, 397-402.

[6] D Joita, Unsupervised Static Discretization Methods in Data Mining, *Revista Mega Byte*, **2010**, 9.

[7] J Pearl, Causality: Models, Reasoning and Inference, *Journal of Econometric Theory*, **2003**, 19, 129-135.

[8] J Seok and YS Kang, Mutual Information between Discrete Variables with Many Categories using Recursive Adaptive Partitioning, *Scientific Reports*, Web. https://www.nature.com/articles/srep10981, **2015**.

[9] https://en.wikipedia.org/wiki/Mutual_information, **2017**.

[10] BC Ross, Mutual Information Between Discrete and Continuous Data Sets, PloS one 9, **2014**.

[11] RG James, JR Mahoney and JP Crutchfield, Trimming the Independent Fat: Sufficient Statistics, Mutual Information, and Predictability from Effective Channel States, *arXiv preprint arXiv:1702.01831*, **2017**.

[12] T Tung, R Taewoo, HL Quang and L Doheon, Inferring Gene Regulatory Networks from Microarray Time Series Data Using Transfer Entropy, *Twentieth IEEE International Symposium on Computer-Based Medical Systems (CBMS'07)*, **2007**, 383-388.

[13] NJ Newton, Transfer Entropy and Directed Information in Gaussian Diffusion Processes, *arXiv preprint arXiv:1604.01969*, **2016**.

[14] K Schindler, M Palus, M Vejmelka and J Bhattacharya, Causality Detection Based on Information-Theoretic Approaches in Time Series Analysis, *Physics Reports*, **2007**, 441(1), 1-46.

[15] https://en.wikipedia.org/wiki/Transfer_entropy, **2017**.

[16] W Zhu, N Zeng and N Wang, Sensitivity, Specificity, Accuracy, Associated Confidence Interval and Roc Analysis with Practical SAS Implementations, *NESUG proceedings: Health Care and Life Sciences*, Baltimore, Maryland, **2010**, 1-9.

[17] D Datyal, A Kaushik and A Tomar, A Novel PCA based Multi-Layer Perceptron Algorithm for Maintainability Prediction, *International Journal of Engineering Trends and Technology*, **2016**, 37(2), 1-7.

[18] M Sokolova and G Lapalme, A Systematic Analysis of Performance Measures for Classification Tasks, *Information Processing & Management*, **2009**, 45(4), 427-437.

[19] Y Sasaki, The Truth of F-measure, *Teach Tutor Mater*, **2007**, 1 (5), 1-5.

[20] T Fawcett, An introduction to ROC analysis, *Pattern Reorganization Letters*, **2006**, 27, 861-874.

[21] https://www.ncbi.hlm.nih.gov/pubmed, **1996**.