2009 International Symposium on Computing, Communication, and Control (ISCCC 2009)
Proc .of CSIT vol.1 (2011) © (2011) IACSIT Press, Singapore

Performance Study on Data Discretization Techniques Using Nutrition Dataset

Nor Liyana Mohd Shuib[1], Azuraliza Abu Bakar[2] and Zulaiha Ali Othman[2+]

[1] Department of Information Science, Faculty of Computer Science and Information Technology, University of Malaya, 50603 Kuala Lumpur, Malaysia

[2] Center for Artificial Intelligence Technology, Faculty of Information Science and Technology, University Kebangsaan Malaysia, 43600 Bangi, Selangor Darul Ehsan, Malaysia

Abstract. Data mining has been widely used in medical and health care domain as the predictive models. Data preprocessing is one of the important steps in data mining process as it consumes about sixty percent of the data mining project effort. Data discretization is one of the pre-processing methods. It makes learning process faster and more accurate. In this paper we proposed the nutrition data classification modeling using two discretization techniques i.e. Boolean Reasoning and Entropy Algorithm. Both techniques are selected from detail study of fifty discretization techniques available to date. The purpose of this work is to compare the performance of different data discretization techniques and to find the most suitable discretization techniques for the nutrition data set. The nutrition data set are obtained from a survey conducted and it contains 160 attributes and 820 records. Both techniques are used to discretize the nutrition data set and the classification performance of both techniques in terms of accuracy and the number of rules evaluated. The experimental results showed that Boolean Reasoning performs better than Entropy Algorithm which gives higher classification accuracy in nutrition data set.

Keywords: Data Mining, Data Pre-processing, Discretization

1. Introduction

There are huge volumes of data available today because of the advancement technology in software computer and media storage. However, these data are often to be dirty due to several reasons such as incomplete data, missing data, noisy data and inconsistent data [1]. The uses of dirty data would give a huge impact to data mining result because it could give wrong interpretation [2]. Therefore, pre-processing is very important to ensure that data to be used are clean and appropriate for data mining. Data preparation and pre-processing is the key to solve the problem [3]. However, pre-processing is always omitted by researcher which will lead to inaccurate model result since the process is time consuming and tedious. A good data pre-processing will helps to create better model and will consume less time.

There are many pre-processing techniques. Each technique has its own functions and advantages. Discretization technique is one of the pre-processing techniques. Most of the classification tasks requires the data to be in the discrete form to be able to perform the mining process. The usage of continuous attributes involves huge storage, misinterpretation and long rules. Hence, discretization is needed to change from continuous attributes to discrete attribute in order to increase the accurateness in prediction.

Discrete attributes are the key factor in data mining as it involves with simple interval numbers for representation which is understandable and easier to use. The rules of discrete attributes usually are shorter and easy to understand, hence will increase the accurateness of prediction. Most of the algorithms in the

[+] Corresponding author. Tel.: 0124591224; fax: 0389216184
E-mail address: aab@ftsm.ukm.my

literature requires discrete attribute which caused data mining practitioners and researchers to perform data discretization before or while doing data mining.

Most of the real data set usually contains continuous attributes. This involves data sets from health care. In this research, nutrition data set from a general hospital in Malaysia which consists of 820 objects and 160 attributes are used. This data set is used to understand the functions of foods and its relation to health. The objective of mining this data set is to identify patients' dietary pattern and how this pattern could lead to the disease. 60 years old people and their dietary pattern have been chosen as the domain. This data set is also used to compare performance of different data discretization techniques and to find the most suitable discretization techniques for the nutrition data set.

2. Literature Review

Data mining has been widely used in medical and health care domain [4; 5; 6; 7]. Data pre-processing is one of the most important steps in data mining process as it consumes about sixty percent of the data mining project effort. The steps are named as data integration, data selection, data cleaning, data reduction and data transformation. Data reduction process refers to two approaches i.e. the reduction of data dimensional sizes or reduction of the data distribution. One of the data distribution reduction approach is data discretization.

Data discretization is defined as one of the way to reduce data used to change the original continuous attributes to discrete attributes [8]. It creates an appropriate number of intervals for data values thus transforming the continuous data values into the discrete values. The smaller data intervals usually contributed to more accurate predictive model which could cover higher prediction rates into new cases. Discretization is required particularly for rule-based data mining model such as decision tree and rough set classifiers.

Based on the study that has been done [9], two types of discretization techniques have been chosen. The methods are Boolean Reasoning [10; 11] and Entropy Algorithm [12].

2.1. Boolean Reasoning

Boolean Reasoning (BR) is suggested by [10]. This technique is used by [11] in rough set theory. BR is developed based on rough set theory and Boolean reasoning. This technique is a supervised technique that consider all attributed at the same time and produce smaller cut point. Cut point is defined as a real value that divides continuous value into intervals [14]. BR was chosen because it is suitable for rough set classification and it was the best approach for researches that involve with classification and recognition [11]. Moreover, this method hasn't been used widely by the researchers in discretization.

2.2. Entropy Algorithm

Entropy Algorithm (Ent-MDLP) is a discretization technique based on entropy that was suggested by Fayyad & Irani [13; 14; 15]. Ent-MDLP uses entropy minimization heuristic (EMH) to discretize continuous attributes to interval. This technique also uses minimum description length criteria [16] to control number of interval. Ent-MDLP is a supervised technique that uses information class entropy to choose cut point. Ent-MDLP method was chosen because it was widely used in discretization researches [13; 17]. It also one of the best discretization methods [18; 19; 20] reported in literature.

3. Model Development

In this research, rough set algorithm is chosen as a data mining tool. Rough set [21] mining algorithm requires the data to be discretized. This is suitable with our research objective which is to compare discretization techniques. This model development can be divided to three steps which is data preparation and data pre-processing, model development and evaluation and testing. The framework for model development is illustrated in Figure 1.

3.1. Preparation of Data and Pre-processing

Data set collected is a nutrition data set from the UKM Hospital (HUKM). These data sets are collected from 820 patients and have 160 attributes. Out of 160 attributes, 56 continuous attributes are identified. First

step in preparing the data is the selection of the attributes. This is done by removing redundant attributes (two attributes that have the same knowledge) and unimportant attributes (attributes that contains insignificant knowledge for modelling). Then, the pre-processing is carried out. The preprocessing stage involves filling in the missing value (use the attribute mean), attribute construction, concept hierarchies, replacement of nearest neighbourhood techniques, and discretization techniques. Techniques Ent-MDLP and BR are chosen for discretization.

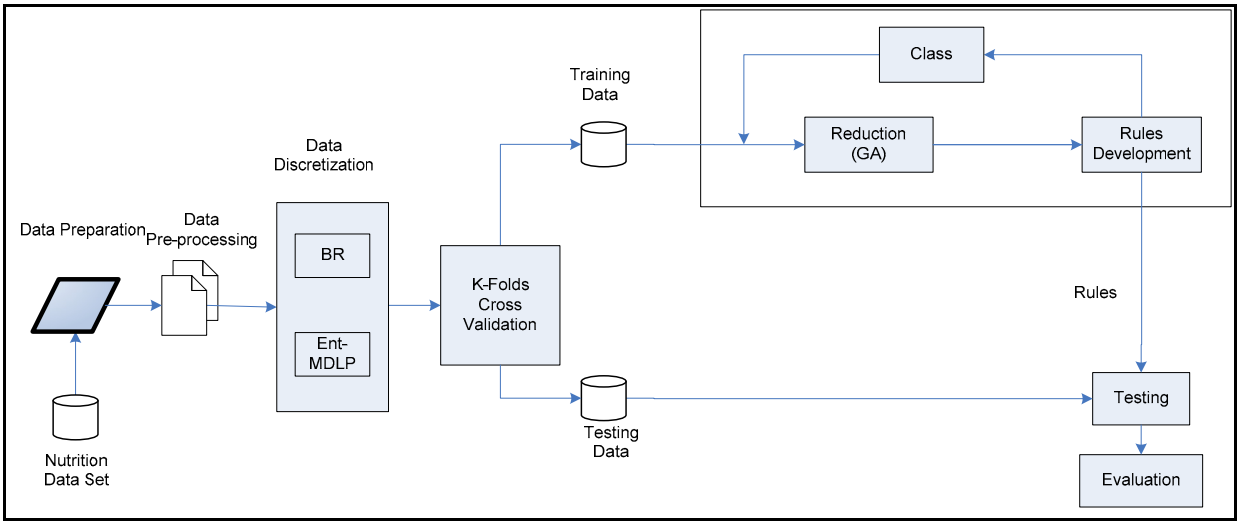

Fig. 1: Model Development Framework

3.2. Mining

After data pre-processing has been conducted, data set is divided to training data and testing data using k-folds cross validation techniques. This technique prepared nutrition data set to 10 folds randomly. Training data is used to develop a model while testing data is used to determine the accuracy of the model acquired. Model development is built using roughs set theory. Two model is developed. One is using data set that is discretized using BR while one data sets i using Ent-MDLP. Rosetta [22], rough set application, is used to mine the data.

3.3. Evaluation and Testing

Evaluation is based on classifier accuracy, numbers of rules, minimum of rules length, maximum of rules length and numbers of intervals.

4. Results

BR and Ent-MDLP has been used as comparison techniques. Evaluation for nutrition data set modeling is based on classifier accuracy (*ACC*), numbers of rules (*NR*), minimum of rules length (*min_L*), maximum of rules length (*max_L*) and numbers of intervals (*I*). the min_L and max_L are considered since it is the indication of complexities of the rules where shorter rules are expected to perform better than longer and specific rules. The best model for all folds are shown in Table 1.

The results showed comparative performance for both BR and Ent-MDLP techniques. However the best accuracy (ACC) was obtained from model 4 using BR recorded 90.52% with 6772 rules. The highest accuracy by Ent-MDLP recorded 87.93% with 7548 rules. In average, BR gives 85.15% accuracy while Ent-MDLP gives 83.27%. Both techniques showed equal performances in min_L and max_L. the number of rules generated (NR) also showed comparative results. For each attribute in the dataset the number of intervals (I) set by both BR and Ent-MDLP showed significant difference. BR outperformed Ent-MDLP by producing the less I. However, this result does not indicate any significant relation to the accuracy for bothtechniques. Naturally, less number of intervals ensured the better modelling accuracy but if the number of intervals are too small it may cause certain information loss in the data. Therefore, in the aspect of number of generated intervals and the quality of rules via knowledge, Ent-MDLP seems to be a better choice.

5. Discussion and Conclusion

- BEST MODEL FROM ALL FOLDS

m	*BR*				*Ent-MDLP*			
	ACC	*NR*	*Min_L*	*Max_L*	*ACC*	*NR*	*Min_L*	*Max_L*
1	84.48	7698	1	3	80.17	6605	1	3
2	85.34	6869	1	3	85.34	6734	1	3
3	86.21	7724	1	3	87.93	7548	1	3
4	90.52	6772	1	3	84.48	6444	1	3
5	87.93	7771	1	3	87.93	7713	1	3
6	78.49	6634	1	3	81.03	7502	1	3
7	81.03	7706	1	3	81.03	7691	1	3
8	84.48	6885	1	3	87.93	7680	1	3
9	83.33	5736	1	3	72.41	7793	1	3
10	89.66	7720	1	3	84.48	7161	1	3

- THE NUMBER OF INTERVALS (I)

Atribut	*prot*	*fat*	*cho*	*ca*	*p*	*fe*	*na*	*k*	*retinol*
BR	3	3	4	2	2	2	5	2	3
Ent-MDLP	336	144	401	165	392	96	380	395	312

In this study, two discretization techniques were used for modelling the patient nutrition data. The experimental results showed that both techniques outperformed in different aspects. BR gives higher accuracy, lesser number of rules and number of intervals. Ent-MDLP gives lower accuracy, larger number of rules and large number of intervals. Both findings have their own advantages and drawbacks. Although BR produced the best model but shorter rules generated may contributes to the loss of knowledge. On the other hand, Ent-MDLP showed comparative performance towards BR with larger number of intervals. It gives good indication that although it produces many distinct values in an attribute which indicate the lesser loss of original knowledge, it does not affects the accuracy of the model.

Discretization techniques are one of the important techniques in data mining. Discrete attributes will produce short and precise results (rules) compared to continuous attributes. This research investigates two types of techniques namely BR and Ent-MDLP to identify the most suitable discretization techniques to the nutrition data sets in order to produce a better model. Based on the experimental results, Boolean Reasoning performs better than Entropy Algorithm which gives higher classification accuracy in nutrition data set. Further research can be made by comparing these techniques in neural network.

6. Acknowledgement

We would like to thank IRPA 04-02-02-004 EA004 group for providing us the nutrition data set to be used in this research.

7. References

[1] A. Storkey. Data Mining and Exploration: Introduction. School of Informatics. 2006. (online) www.inf.ed.ac.uk/teaching/ courses/dme/slides/intro4up.pdf.

[2] P. Wright. Knowledge discovery preprocessing: determining record usability. *ACM Southeast Regional Conference*. 1998, pp. 283-288.

[3] D. Pyle. *Data preparation for data mining*. San Francisco: Morgan Kaufmann Publishers,1999.

[4] A. Kusiak, K.H. Kernstine, J.A. Kern, K.A. McLaughlin, and T.L. Tseng. Data Mining: Medical and Engineering Case Studies. *Proceedings of the IIE Research 2000 Conference*, Cleveland, OH, May 2000, pp. 1-7

[5] I. Kononenko, I. Bratko, and M. Kokar, Application of machine learning to medical diagnosis, in

[6] Michalski, RS, Bratko, I and Kubat M. (Eds), Machine Learning in Data Mining: Methods and

[7] Applications, Wiley, New York, 1998, pp. 389-428.

[8] M.R. Kraft, K. C Desouza and I. Androwich. Data Mining in Healthcare Information Systems: Case Study of a Veterans' Administration Spinal Cord Injury Population. In *Proceedings of the 36th Annual Hawaii international Conference on System Sciences (Hicss'03).* IEEE Computer Society, Washington. 2003, **6** (6): 159.1.

[9] Peng Liu, Lei Lei, Junjie Yin, Wei Zhang, Wu Naijun and E. El-Darzi. Healthcare Data Mining: Prediction Inpatient Length of Stay. *3rd International IEEE Conference on Intelligent Systems.* 2006, pp 832-837.

[10] N. Goharian and D. Grossman. Data mining: data preprocessing. Illinois Institute of Technology. 2003. (online) www.ir.iit.edu/~nazli/cs422/ CS422-Slides/DM-Preprocessing.pdf

[11] Nor Liyana Mohd Shuib. Discretization Techniques in Data Mining: A Case Study on Nutrition Data Set. Master Thesis. Universiti Kebangsaan Malaysia. 2008.

[12] H.S. Nguyen, and A. Skowron. Boolean reasoning for feature extraction problems. *International Symposium on Methodologies for Intelligent System.* 1997, pp.117-126.

[13] Z. Pawlak and A. Skowron. Rough sets and Boolean reasoning. *Information Sciences*. 2007. 177(1): pp. 41-73.

[14] U.M. Fayyad and K.B. Irani, Multi-interval discretization of continuous valued attributes for classification learning. *Proceedings of IJCAI 2*. 1993, pp.1022-1027.

[15] J. Dougherty, R. Kohavi and M. Sahami. Supervised and unsupervised discretization of continuous features. *Proc Twelfth International Conference on Machine Learning*. 1995, pp.194-202.

[16] H. Liu, F. Hussain, C.L Tan and M. Dash. Discretization: an enabling technique. *Data Mining and Knowledge Discovery.* 2002, **6**: pp.393-423.

[17] Ying Yang. Discretization for Naive-Bayes Learning. Tesis Ph.D. Monash University. 2003.

[18] J. Rissanen. Modelling by shortest data description. *Automatica*. 1978, **14**: pp.465-471.

[19] J. Gama and C. Pinto. Discretization from data streams: applications to histograms and data mining. *Proceedings of the 2006 ACM symposium on Applied computing*. 2006, pp. 662 – 667.

[20] J. Cerquides and R López de Màntaras. Proposal and empirical comparison of a parallelizable distance-based discretization method. *3d Int. Conference on Knowledge Discovery and Data Mining (KDD'97).* 1997. pp.139-142.

[21] X. Liu and H. Wang. A discretization algorithm based on a heterogeneity criterion. *IEEE Transactions on Knowledge and Data Engineering*. 2005, **17**(9): pp.1166-1173.

[22] F. Tay and L. Shen. A modified chi2 algorithm for discretization. *IEEE Transactions of Knowledge and Data Engineering*. 2002, **14**(3): pp.666-670.

[23] Z. Pawlak. Rough sets. *International Journal of Computer and Information Science.* 1982, **11**: pp.341-356.

[24] A. Øhrn. ROSETTA: Technical Reference Manual. 1999. (online)

[25] http:// www.idi.ntnu.no/¸aleks/rosetta.