A comparative analysis of discretization methods for Medical Datamining with Naïve Bayesian classifier

Ranjit Abraham,
Dept. of Computer Science, TocH Institute of Sci. and Tech., Arakkunnam, Kerala, INDIA.
ranjit.abraham@gmail.com

Jay B.Simha,
ABIBA Systems, Bangalore, INDIA.
jay.b.simha@abibasystems.com

Iyengar S.S
Dept. of Computer Science, Louisiana State University, Baton Rouge, USA.
iyengar@bit.cse.lsu.edu

Abstract

Naïve Bayes classifier has gained wide popularity as a probability-based classification method despite its assumption that attributes are conditionally mutually independent given the class label. This paper makes a study into discretization techniques to improve the classification accuracy of Naïve Bayes with respect to medical datasets. Our experimental results suggest that on an average, with Minimum Description Length (MDL) discretization the Naïve Bayes Classifier seems to be the best performer compared to popular variants of Naïve Bayes as well as some popular non-Naïve Bayes statistical classifiers.

1. Introduction

With the widespread use of computers in medical practice several computer programs have been developed to carry out optimal management of data for extraction of knowledge or patterns contained in the data. These include Expert Systems, Artificial Intelligence and Decision Support Systems. One such program approach has been data classification using naïve Bayes, which has gained much prominence because of its simplicity and comparable accuracy with other classifiers. In this paper, we show that it is possible to reliably improve the naïve Bayes classifier by using data discretization as part of data pre-processing.

2. Naïve Bayes and NB Classifier

Naïve Bayes (NB), a special form of Bayesian Network has been widely used for data classification in that its predictive performance is competitive with state-of-the-art classifiers [1]. As a classifier, it learns from training data from the conditional probability of each attribute given the class label. It uses Bayes rule to compute the probability of the classes given the particular instance of the attributes, prediction of the class is done by identifying the class with the highest posterior probability. Research shows naïve Bayes still performs well in spite of strong dependencies among attributes.

The naïve Bayesian classifier represented as a Bayesian network has the simplest structure. The assumption made is that all attributes are independent given the class and takes the form

$$c(E) = \arg\max_{c \in C} \; p(c) \prod_{i=1}^{n} p(x_i \mid c)$$

where x_i is the value of the attribute X_i and c the class value for the class variable C.

3. Discretization for NB Classifier

Research study shows that naïve Bayes classification works best for discretized attributes and discretization effectively approximates a continuous variable [2].

Both Equal Width (EWD) and Equal Frequency (EFD) discretization are unsupervised direct methods and have been used because of their simplicity and reasonable effectiveness [2]. Both EWD and EFD suffer from possible attribute loss on account of the pre-determined value of intervals k.

The Minimum Description Length (MDL) discretization is entropy based heuristic given by Fayyad and Irani [3]. The technique evaluates a candidate cut point between each successive

9th International Conference on Information Technology (ICIT'06)
0-7695-2635-7/06 $20.00 © 2006 IEEE

pair of sorted values.

4. Experimental Evaluation

We have used 10-fold cross validation test method to 28 of the publicly available medical datasets. Discretization techniques were employed as pre-processing to the datasets. In both EWD and EFD ten bins were assumed. The win-lose-tie given at the bottom of Table 1 show that Fayyad and Irani's MDL discretization, on the average improved classification accuracy compared to that of EWD and EFD discretization. Table 2 shows the accuracy results for different classifiers, which include variants of naïve Bayes, and popular non-Naïve Bayes statistical classifiers [4]. The wins given at the bottom of Table 2 indicate the superiority of NB with MDL discretization for our experiments. We argue that MDL discretization does better on account of using the class information entropy after discretization and, EWD and EFD discretization levels are not optimized. This research augments the argument simple methods are better in medical data mining.

5. Conclusions

In this research work an attempt was made to evaluate the naïve Bayes classifier that could be used for medical datamining. Our experimental results indicate that, on an average, naïve Minimum Description Length (MDL) discretization seems to be the best performer compared to the considered various naïve Bayes and non-Naïve Bayes classifiers. The work is presently under progress to explore feature selection methods in achieving better naïve Bayes classification performance.

6. References

[1] Duda and Hart. "Pattern Classification and Scene Analysis" 1973, John Wiley and Sons, NY.

[2] Chun-Nan Hsu, Hung-Ju Huang and Tsu- Tsung Wong. "Why Discretization works for Naïve Bayesian Classifiers", *17th ICML*, 2000, pp 309-406.

[3] Fayyad U. M. and Irani K. B., "Multi-interval discretization of continuous-valued attributes for classification learning". *In Proceedings of the 13th International Joint Conference on Artificial Intelligence*, 1993, pp. 1022–1027.

[4] Ranjit Abraham, Simha J.B., Iyengar S.S., "Medical data mining with probabilistic classifiers", Working paper, Department of Computer Science, Louisiana State University, Baton Rouge, USA, 2006.

Table 1: Naïve Bayes classification accuracy with and without discretization

Sl. No.	Medical Dataset	NB without Discretization	Naïve Bayes with Discretization		
			EWD	EFD	MDL
1	Wisconsin Breast Cancer (699, 10, 9, 2, Yes, No)	95.9943	97.2818	97.2818	96.9957
2	Cleveland Heart Disease (303, 14, 5, 2, Yes, No)	83.8284	83.4983	83.4983	83.8284
3	Hepatitis (155, 20, 6, 2, Yes, No)	84.5161	84.5161	83.2258	84.5161
4	Thyroid -new (215, 6, 5, 3, No, No)	96.7442	92.093	95.814	96.2791
5	Appendicitis (106, 9, 8, 2, Yes, No)	84.9057	81.1321	84.9057	88.6792
6	Pancreatic Ca biomarkers (141, 3, 2, 2, No, No)	73.7589	63.1206	74.4681	78.7234
Win-Lose-Tie with respect to Naïve Bayes classification (without discretization)			1- 4- 1	2- 3- 1	3- 2- 1

(Within Brackets are given total instances of the dataset, total number of attributes, total number of attributes discretized, number of classes, status of missing attribute values and status of noisy attribute values).

Table 2: Classification accuracy of Naïve Bayes, variants of NB and non-NB classifiers

Sl. No.	Medical Dataset	Classifier Accuracy								
		NB	NB (MDL)	Variants of NB				Popular Classifiers		
				SNB	BNB	TAN	FAN	DT	k-NN	LR
1	Wisconsin Breast Cancer	95.9943	96.9957	96.1373	95.5651	96.7096	95.5651	94.5637	95.279	96.9665
2	Cleveland Heart Disease	83.8284	83.8284	84.4884	83.4983	83.4983	81.5182	75.9076	76.8977	84.8185
3	Hepatitis	84.5161	84.5161	84.5161	85.8065	83.2258	83.2258	83.871	80.6452	82.5806
4	Thyroid –new	96.7442	96.2791	97.6744	95.814	94.4186	95.3488	92.093	97.2093	96.7442
5	Appendicitis	84.9057	88.6792	80.1887	83.0189	87.7358	86.7925	86.7925	83.0189	87.7358
6	Pancreatic Ca biomarkers	73.7589	78.7234	73.7589	73.7589	70.992	72.3404	73.0496	72.3404	80.1418
Wins		0/6	2/6	1/6	1/6	0/6	0/6	0/6	0/6	2/6

Abbreviations: NB- Naïve Bayes**, NB (MDL) –** Naïve Bayes with MDL discretization, **SNB –** Selective Naïve Bayes**, BNB-** Boosted Naïve Bayes, **TAN**- Tree Augmented Naïve Bayes, **FAN** – Forest Augmented Naïve Bayes, **DT** – Decision Tree, **k-NN**- k -Nearest Neighbor, **LR**- Logistic Regression