A COMPARATIVE ANALYSIS OF BAYESIAN METHODS FOR REAL ESTATE DOMAIN

Geetali Banerji*

Kanak Saxena**

Abstract

Bayesian 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 various algorithms to improve the classification accuracy of Bayesian methods with respect to real estate datasets. We have applied Bayesian methods on two variations of data sets in three different test modes. In the first instance we have taken complete data sets, our experimental results suggest that, Bayesian network Classifier seems to be the best performer compared to popular variants of Bayesian classifiers. In second instance we have applied the same techniques on selected attribute i.e. after removing demographic details of customers and and found that there is a drastic change in the results of various Bayesian techniques except Complement Naive Bayes which is giving near about same accuracy and error rate in both variations i.e. it is unaffected with the attribute sets.

Keywords- Classification, Naïve bayes, Bayesian Network, Complement Naïve Bayes

* Information Technology Department, Institute of Information Technology & Management, New Delhi, India.

** Department of Computer Applications, Samrat Ashok Technological Institute, Vidisha, M.P., India.

I INTRODUCTION

Classification is a basic task in data analysis and pattern recognition that requires the construction of a *classifier*, that is, a function that assigns a *class* label to instances described by a set of *attributes*. The induction of classifiers from data sets of preclassified instances is a central problem in machine learning. Numerous approaches to this problem are based on various functional representations such as decision trees, decision lists, neural networks, decision graphs, and rules. One of the most effective classifiers, in the sense that its predictive performance is competitive with state-of-the-art classifiers, is the so-called *naive Bayesian* classifier [10].

II BAYESIAN METHODS

Bayesian Network

A Bayesian Network (BN) is a graphical model for probability relationships among a set of variables features. The Bayesian network structure S is a directed acyclic graph (DAG) and the nodes in S are in one-to-one correspondence with the features X. The arcs represent casual influences among the features while the *lack* of possible arcs in S encodes conditional independencies. Moreover, a feature (node) is conditionally independent from its non-descendants given its parents ($X1$ is conditionally independent from $X2$ given $X3$ if P ($X1|X2$, $X3$) =P ($X1|X3$) for all possible values of $X1$, $X2$, $X3$).

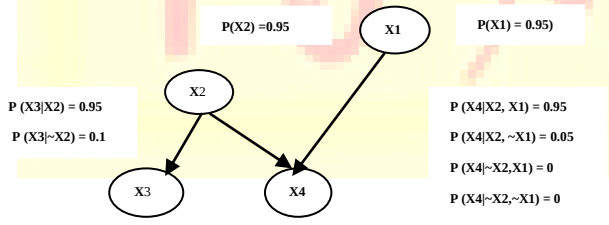


Fig.1 The structure of Bayes network

Typically, the task of learning a Bayesian network can be divided into two subtasks: initially, the learning of the DAG structure of the network, and then the determination of its parameters. Probabilistic parameters are encoded into a set of tables, one for each variable, in the form of local conditional distributions of a variable given its parents. Given the independences encoded into the network, the joint distribution can be reconstructed by simply multiplying these tables. Within the general framework of inducing Bayesian networks, there are two scenarios: known structure and unknown structure. In the first scenario, the structure of the network is given (e.g. by an expert) and assumed to be correct. Once the network structure is fixed, learning the parameters in the Conditional Probability Tables (CPT) is usually solved by estimating a locally exponential number of parameters from the data provided [6]. Each node in the network has an associated CPT that describes the conditional probability distribution of that node given the different values of its parents. In spite of the remarkable power of Bayesian Networks, they have an inherent limitation. This is the computational difficulty of exploring a previously unknown network. Given a problem described by *n* features, the number of possible structure hypotheses is more than exponential in *n*. If the structure is unknown, one approach is to introduce a scoring function (or a score) that evaluates the "fitness" of networks with respect to the training data, and then to search for the best network according to this score. Several researchers have shown experimentally that the selection of a single good hypothesis using greedy search often yields accurate predictions [7]. The most interesting feature of BNs, compared to decision trees or neural networks, is most certainly the possibility of taking into account prior information about a given problem, in terms of structural relationships among its features. This prior expertise, or domain knowledge, about the structure of a Bayesian network can take the following forms:

1. Declaring that a node is a root node, i.e., it has no parents.

2. Declaring that a node is a leaf node, i.e., it has no children.

3. Declaring that a node is a direct cause or direct effect of another node.

4. Declaring that a node is not directly connected to another node.

5. Declaring that two nodes are independent, given a condition-set.

6. Providing partial nodes ordering, that is, declare that a node appears earlier than another node in the ordering.

7. Providing a complete node ordering.

A problem of BN classifiers is that they are not suitable for datasets with many features [18]. The reason for this is that trying to construct a very large network is simply not feasible in terms of time and space. A final problem is that before the induction, the numerical features need to be discredited in most cases. [5]

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

$$c(E) = \arg\max_{c \in C} p(c) \prod_{i=1}^{n} p(xi|c)$$

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

Complement Naïve Bayes (CNB)

The CNB is a method that tackles the ununiformity of the data distribution. The CNB classifier is a modification of the NB classifier. This classifier improves classification accuracy by using data

from all categories except the category which is focused on [17]. This classifier is also used as a baseline.

III EXPERIMENTAL EVALUATION

Analysis

We have used 10-fold cross validation test using WEKA version 3-6-2[8, 13, 14, and 15] (Fig. 2) to 5821 Real estate datasets. The following are the two categories under which the analysis are carried out and the factors on which the analysis are carried out on Kappa Statistic, Mean absolute error, Root Mean Squared Error, Root Absolute error and Root Relative Square Error and Execution Time.[9]

- Considering complete attribute set for applying various methods.
- Methods applied on selected attributes (Demographic details removed)

Each of the method is tested on following 3 different test modes:

1. Cross validation with 10 fold

2. Splitting Data Set (75% Training and 25% Test data set)

3. Complete data set as Training data set

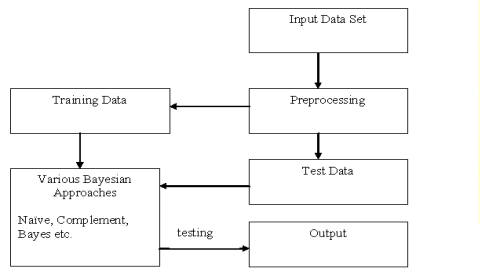


Fig. 2 Weka Model

Cohen's Kappa: Measures of Data Consistency

Cohen's kappa measures the agreement internal consistency based on a contingency table. In this context a measure of agreement assesses the extent to which two raters give the same ratings to the same objects. The set of possible values for one rater forms the columns and the same set of possible values for some second rater forms the rows.

Kappa κ = [observed concordance - concordance by chance]/ [1- concordance by chance]

Where "by chance" is calculated as in chi-square: multiply row marginal times column marginal and divide by n. One may use this measure as a decision-making tool:

Kappa κ	Interpretation
$\kappa < 0.00$	Poor
$0.00 \leq \kappa < 0.20$	Slight
$0.20 \leq \kappa < 0.40$	Fair
$0.40 \leq \kappa < 0.60$	Moderate
$0.60 \leq \kappa < 0.80$	Substantial
$0.80 \leq \kappa$	Almost Perfect

This interpretation is widely accepted, and many scientific journals routinely publish papers using this interpretation for the result of test of hypothesis. [12]

Mean absolute error

Let D is dataset with values (x_1, y_1) , $(x_2, y_2),\ldots, (x_d, y_d)$. Let y_i is the actual value and y'_i is the predicted value for the independent variable x_i . y is the mean value of y_i. Mean absolute error is defined as $Mean\ absolute\ error = \frac{\sum_{i=1}^{d} |yi - yi'|}{d}$

$Mean\ Squared\ error = \frac{\sum_{i=1}^{d} (yi - yi')2}{d}$

Square root of the mean squared error is called as root mean squared error.

$$Relative\ absolute\ error = \frac{\sum_{i=1}^{d}|yi - yi'|}{\sum_{i=1}^{d}|yi - y|}$$

$$Root\ Relative\ Squared\ error = \sqrt{\frac{\sum_{i=1}^{d}(yi - yi')2}{\sum_{i=1}^{d}(yi - y)2}}$$

The mean squared error exaggerates the presence of outliers, while the mean absolute error does not. If we were to take the square root of the mean squared error, the resulting error measure is called the root mean squared error. This is useful in that it allows the error measured to be of the same magnitude as the quantity being predicted. In practice, the choice of error measure does not greatly affect prediction model selection but it reflects the deviation from the actual values.

Time taken

Time taken to perform the algorithm is also important in terms of computation complexity of the algorithm on the machine.

Findings

Table 1 shows the results of various Bayesian algorithms on real estate complete data set on 3 test modes. It is found that Bayesian Network is best among all other methods in terms of identifying correct instances and low error rates. Training Test mode is best among three test modes because it considers all the possibilities during testing.

Table 2 shows the results of various Bayesian algorithms on Real Estate selected data set (excluding demographic details) on 3 test modes. It is found that Bayesian Network is best among all other methods in terms of identifying correct instances and low error rates. Training Test mode is best among three test modes.

Graph 1 show the outcome of various Bayesian algorithms on selected attributes (Demographic details are removed). Graph 2 depicts the outcome of various Bayesian algorithms on complete attribute set. The comparison between both variations are depicted in Graph 3, it is very clear

that all the methods perform well in case of complete attribute set except Complement Naïve Bayes which outperforms in case of selected attribute set.

IV CONCLUSIONS

In this research work an attempt was made to evaluate the naïve Bayes classifier that could be used for real estate data sets. Our experimental results indicate that, the Bayesian network seems to be the best performer compared to the considered various naïve Bayes classifiers on selected as well as complete data set of Real Estate. It is proved that Complement Naïve Bayes performs well on an average same in case of complete/selected dataset. The results are improved after removing the demographic details of the customer. But In context to India, these factors are very important in identifying the purchasing behavior of a customer.

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.

[5] Thair Nu Phyu, "Survey of Classification Techniques in Data Mining ", Proceedings of the International MultiConference of Engineers and Computer Scientists 2009 Vol I IMECS 2009, March 18 - 20, 2009, Hong Kong

[6] Jensen, F. (1996). "An Introduction to Bayesian Networks". Springer.

[7] Chickering D. M. & D. Heckerman (1996), "Efficient approximations for the marginal likelihood of incomplete data given a Bayesian network", In E. Horvits & F. Jensen (Eds.), Proceedings of the Twelfth Conference on Uncertainty in Artificial Intelligence (pp. 158–168). San Francisco, CA: Morgan Kaufmann

[8]R. Bouckaert Remco , Eibe Frank et. al, "WEKA Manual for Version 3-6-2",January 11, 2010

[9]Geetali Banerji, Kanak Saxena, "Predictive Model- A Boon for real estate" , International Journal for Wisdom Based Computing Volume(1) 2, April 2012

[10] Geetali Banerji, Kanak Saxena," An Algorithm for Rule based Classification", Emerging Trends in Information Technology 2012, IITM, New Delhi

[11] Geetali Banerji, Kanak Saxena,"Analysis of Data Mining techniques on Real Estate", International Journal of Soft Computing and Engineering (IJSCE) ISSN: 2231-2307 Volume-2, Issue-3, July 2012,(Under Publication)

[12] http://home,ubalt.edu/ntsbarsh/stat-data

[13] Witten I. frank, E. (2000),"Data Mining: *practical machine learning tools and techniques with Java implementations, Morgan Kaufmann*", San Francisco.

[14] CBA. Data mining tool. Downloading page. http://www.comp.nus.edu.sg/~dm/p_download.html. Viewed on February 2010.

[15] Weka. Data Mining software in Java. http://www.cs.waikato.ac.nz/ml/weks. Viewed on February 2010

[16] J.D.M.Rennie, L.Shih, J.Teevan, and D.R.Karger, 2003. "Tackling the poor assumptions of naive bayes text classification." In ICML2003, pages 616–623.

[17] Kanako Komiya, Naoto Sato er. Al., "Negation Naïve Bayes for Categorization of Product Pages on the Web", Proceedings of recent advances in Natural Language Processing, pages 586-591, Hissar, Bulgaria, 12-14 September 2011

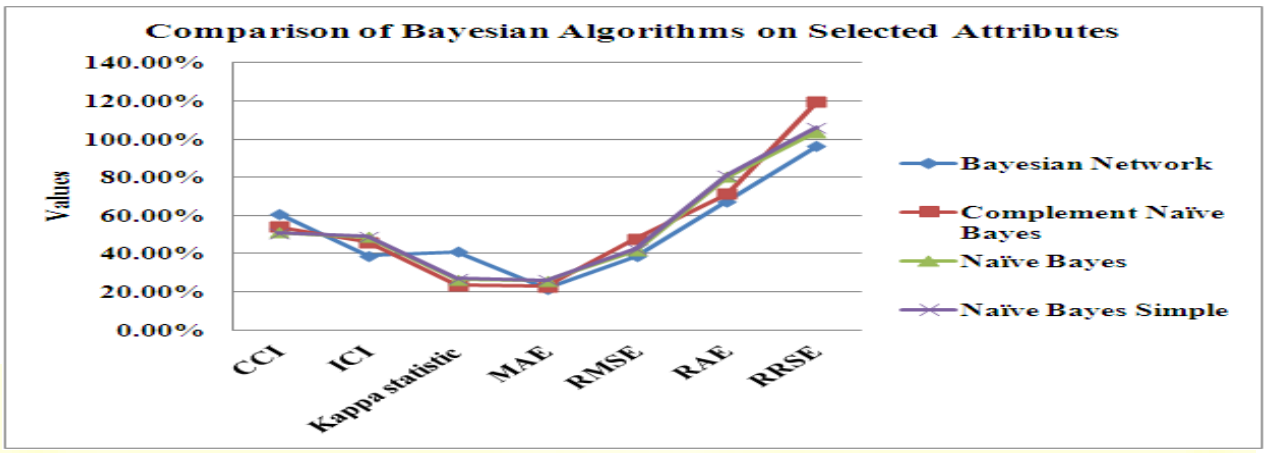
[18] Cheng J. Greiner, R, (2001). "Learning Bayesian Belief Networks Classifiers: Algorithms and Systems, In Stroulia, E. & Marwin, S.(ed.), AI 2001,, 141-151, LNAI 2056

Table 1. Comparison of various Bayesian algorithms (3 test modes) on Real Estate Complete Data set

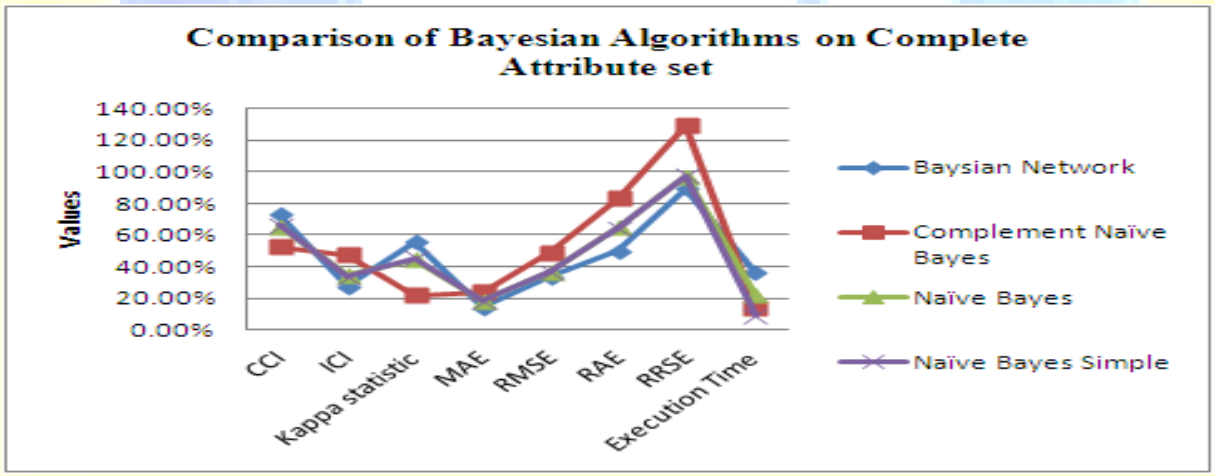
Measuring Modes*	Bayesian Network			Complement Naïve Bayes			Naive Bayes			Naive Bayes Simple		
	1	2	3	1	2	3	1	2	3	1	2	3
CCI (%)	72.70	72.65	73.66	52.38	51.66	52.59	65.50	65.98	66.60	66.29	66.39	67.27
ICI (%)	27.30	27.35	26.34	47.62	48.34	47.41	34.50	34.02	33.40	33.71	33.61	32.73
Kappa Statistic	0.555	0.5629	0.572	0.2195	.2076	.2233	0.4424	.4595	.4612	0.453	.4688	.4691
MAE	0.143	0.1419	0.1389	0.2381	.2417	.2371	0.1855	.1843	.1818	0.1839	.1851	.1805
RMSE	0.3377	0.3345	0.3319	0.488	.4916	.4869	0.3679	0.3661	.3632	0.3691	.3702	.3647
RAE (%)	49.99	49.39	48.46	83.09	84.35	82.73	64.75	64.17	63.43	64.17	64.43	63.00
RRSE (%)	89.21	88.07	87.71	128.93	129.90	128.65	97.20	96.38	95.95	97.51	97.46	96.36
Execution Time(in secs)	0.36	0.3	0.42	0.14	.03	0.5	0.22	0.2	0.2	0.09	.08	.011

Table 2. Comparison of various Bayesian algorithms (3 test modes) on Real Estate Selected Data set

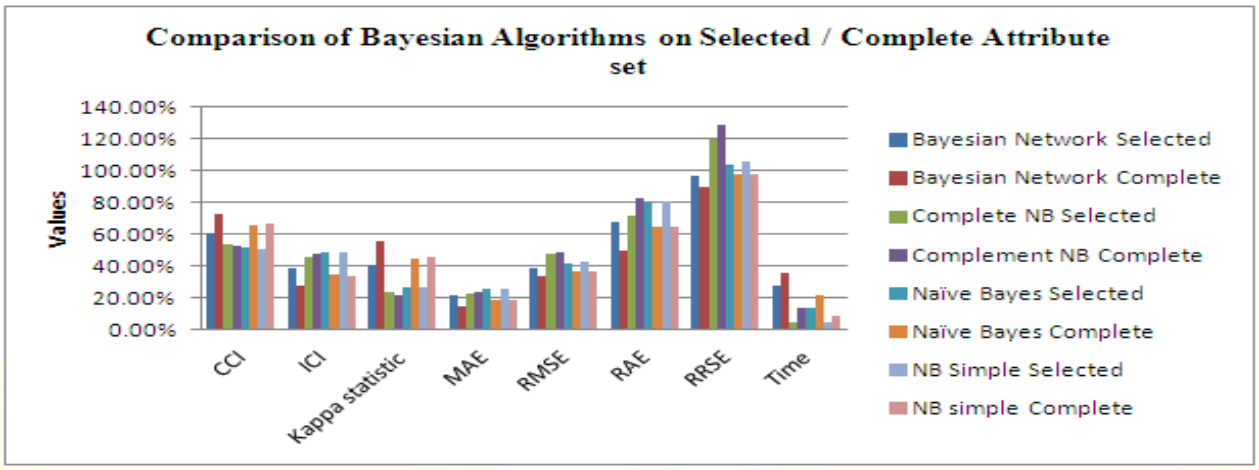
Measuring Modes*	Bayesian Network			Complement Naïve Bayes			Naive Bayes			Naive Bayes Simple		
	1	2	3	1	2	3	1	2	3	1	2	3
CCI (%)	61.06	74.43	74.95	54.04	51.96	51.66	51.23	66.94	67.17	50.87	67.15	67.31
ICI (%)	38.94	25.57	25.05	45.96	48.041	48.34	48.77	33.06	32.83	49.13	32.85	32.69
Kappa Statistic	0.4102	.5785	0.581	0.236	.2157	.2076	0.2644	0.46	.4596	0.2693	.4635	0.463
MAE	0.2167	.1408	.1382	0.2298	.2402	.2417	0.2585	.1845	.1848	0.2608	.1831	.1834
RMSE	0.3861	.3197	.319	0.4794	.4901	.4916	0.4165	.3523	.3531	0.4253	.3534	.3541
RAE (%)	67.40	49.00	48.22	71.46	83.63	84.35	80.40	64.22	64.50	81.12	63.75	64.02
RRSE (%)	96.29	84.16	84.27	119.55	129.06	129.90	103.87	92.75	93.29	106.08	93.039	93.55
Execution Time(in secs)	0.28	.22	.2	0.05	.03	.03	0.14	.13	.34	.05	0.3	.08



Graph 1. Bayesian Algorithms on Selected Attribute



Graph 2. Bayesian Algorithms on Complete Attribute set



Graph 3. Bayesian Algorithms on Selected /Complete Attribute set

DATADICTIONARY

NO.	Name	Description	NO.	Name	Description
1	CUSTYPE	Customer Subtype	21	SRS	Service
2	NOH	Number of houses	22	MGMT	Management
3	ASH	Avg size household	23	TL	Trained labor
4	AS	Avg age	24	UTL	Untrained labor
5	CMT	Customer main type	25	SCA	Social class A
6	NAT	Nationality	26	SCB1	Social class B1
7	CS	Caste	27	SCB2	Social class B2
8	SCS	Subcaste	28	SCC	Social class C
9	NR	No religion	29	SCD	1Social class D
10	MRD	Married	30	RH	Rented house
11	LT	Living together	31	HO	Home owners
12	OR	Other relation	32	C1	1 car
13	SNG	Singles	33	C2	2 cars
14	HWTC	Household without children	34	NC	No car
15	HWTC	Household with children	35	PL	Policy Investment
16	HLE	High level education	36	OI	Other Investment
17	MLE	Medium level education	37	INL	Income < 20.000
18	LLL	Lower level education	38	INM	Income 20-55.000
19	HS	High status	39	INMH	Income 55-85.000
20	BUS	Business	40	INH	Income 85-125.000
41	INHH	Income >123.000	42	AI	Average income
43	IPC	Investment power class			

***CCI: Correctly Classified Instances, ICI: Incorrectly Classified Instances, MAE: Mean Absolute Error, RMSE: Root Mean Squared Error, RAE: Relative Absolute Error (%), RRSE: Root Relative Squared Error (%), Time: Execution Time (in seconds)**