Environmental Prediction System Using Naïve Bayes' Classifier

Rupali Borase[1], Prof. Priyanka Kedar[2]
[1, 2] Dept of Computer Engineering
[1, 2] D.P.C.O.E. Pune.

Abstract- *Feature selection is a key pre-preparing procedure for choosing more pertinent highlights and destroying the repetitive qualities. Finding the more important highlights for the objective is a fundamental movement to enhance the prescient exactness of the learning calculations since more superfluous highlights in the first element space will cause more order mistakes and expend more opportunity for learning. Numerous techniques have been proposed for highlight significance investigation however no work has been finished utilizing Bayes Theorem, Thus this venture has been started to present a class of play and no play likelihood by utilizing Naïve bayes arrangement. The principle goal of acquainting this approach is with improve the prescient precision of the Naive Bayesian Classifier.*

Keywords- Bayes Theorem, Naive Bayesian Classifier, Feature Selection, Irrelevant and Redundant Attributes.

I. INTRODUCTION

In Feature Selection (FS) is a viable pre-preparing procedure ordinarily utilized as a part of information mining, machine learning and counterfeit consciousness in lessening dimensionality, expelling unessential and excess information, expanding learning precision and decreasing the pointless increment of computational cost [10]. Extensive number of highlights given as contribution to the arrangement calculations may prompt inadequate memory and furthermore require more opportunity for learning. The highlights which don't have any effect on the objective is said to be immaterial. The immaterial highlights display in the first component space will deliver more order mistakes and some of the time may create much more dreadful outcomes. In this way it is basic to choose the more applicable highlights which will give helpful data to the objective and it can be performed through FS. Highlight pertinence is ordered into three classifications viz., emphatically important, feebly pertinent and unimportant. An emphatically significant element is constantly fundamental for the ideal subset and it can't be expelled. A component is said to be feebly applicable on the off chance that it is fundamental for an ideal subset just at specific conditions. An insignificant element is one which isn't vital at all and henceforth it must be evacuated. Hence an ideal subset of highlights ought to incorporate all emphatically important, a subset of feebly applicable.

The target of the proposed work is to upgrade the prescient precision of the Naive Bayesian Classifier (NBC) with restricted subset of chose highlights. The NBC is a measurable classifier in light of BT. since the Bayesian investigation experiences high computational cost particularly in models with an extensive number of highlights and to diminish the model development time of NBC, this work utilizes BT in the pre-handling advance in finding the more important highlights.

II. MATHEMATICAL PRELIMINARIES

Bayes Theorem

This hypothesis is named after Thomas Bayes (/ˈbeɪz/or "coves") and regularly called Bayes' law or Bayes' rule Formula

The equation used is:

$$P(X)=P(X|Y)P(Y|X)/P(X)$$

Where:

P(X) prior probability or marginal probability of X. It is "prior" in the sense that it does not take into account any information about Y.

P(X|Y) conditional probability of X, given Y = posterior probability because it is derived from or depends upon the specified value of Y.

P(Y|X) conditional probability of Y given X = likelihood.
P(Y) prior or marginal probability of Y, and acts as a normalizing constant

Naive Bayesian classifier

It is a measurable classifier in light of the Bayes Theorem. Give D a chance to be an arrangement of preparing tuples with class name A. Assume there are m particular classes A1,A2,… ,Am. The part of this classifier is to anticipate that the given tuple B has a place with the class having the most elevated back likelihood contained on C. i.e., the tuple C has a place with Ai iff for $1 \leq q \leq r$ and $q \neq p$, is registered as [3] P (Ap / B)> P(Cq/C) for $1 \leq q \leq r$ and $q \neq p$, P(Ap/B) is registered as [3].

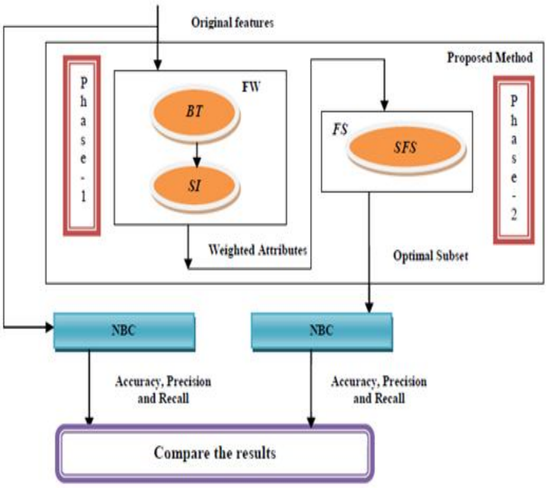

Fig 1: Framework of Proposed Work

To show the relevance of the proposed work, the weather dataset has been taken from UCI machine learning repository. The dataset comprises 5 fields, out of which 2 are continuous and 3 are discrete. The dataset contains 14 instances. The target attribute contains two distinct values 'play' and 'not play'. The entire content of the weather dataset is shown in table 1.

Table No.1 Play Cricket - Training Example

Outlook	Temperature	Humidity	Windy	Class
S	H	P	F	N
S	H	P	T	N
O	H	P	F	P
R	M	P	F	P
R	C	N	F	P
R	C	N	T	N
O	C	N	T	P
S	M	P	F	N
S	C	N	F	P
R	M	N	F	P
S	M	N	T	P
O	M	P	T	P
O	H	N	F	P
R	M	P	T	N

Outlook:
Sunny=S Overcast=O Rain=R
Temperature: Hot=H Mild=M Cool=C
Humidity: Peak=P Normal=N
Windy: True=T False=F
P(Play=Yes) = 9/14
P(Play=No) = 5/14

Table No.2 Probability of Play and No Play for Outlook

Outlook	Play=Yes	Play=No
Sunny=S	2/9	3/5
Overcast=O	4/9	0
Rain=R	3/9	2/5

Table No. 3 Probability of Play and No Play for Temp.

Temperature	Play=Yes	Play=No
Hot=H	2/9	2/5
Mild=M	4/9	2/5
Cool=C	3/9	1/5

Table No.4 Probability of Play and No Play for Humidity

Humidity	Play=Yes	Play=No
Peak=P	3/9	4/5
Normal=N	06-May	1/5

Table No.5 Probability of Play and No Play for Windy

Windy	Play=Yes	Play=No
True=T	2/9	1/5
False=F	6/9	2/5

Example for Not Play:

Test Phase, Given a new instance, predicts its label
x=(Outlook=Sunny, Temperature=High, Humidity=Peak, Windy=False)
– Look up tables achieved in the learning phrase
– Decision making
P(Outlook=Sunny|Play=No) = 3/5
P(Temperature Hot|Play=No) =2/5
P(Humidity=Peak|Play=No) = 4/5
P(Windy=False|Play=No) = 2/5
P(Play=No) = 5/14
P(Outlook=Sunny|Play=Yes) = 2/9
P(Temperature=Hot|Play=Yes) = 2/9
P(Humidity=Peak|Play=Yes) = 3/9
P(Windy= False||Play=Yes) = 6/9
P(Play=Yes) = 9/14
P(Yes|x)≈
[P(Sunny|Yes)P(Hot|Yes)P(Peak|Yes)P(False|Yes)]P(Play=Yes) = 0.00705467
P(No|x)≈[P(Sunny|No)
P(Hot|No)P(Peak|No)P(False|No)]P(Play=No) = 0.02742857
Given the fact P(Yes|x) < P(No|x), we label x to be “No”.

The new tuple classified under "NOT PLAY" category. Hence there will be NOT play.

Example for Play:

Test Phase, Given a new instance, predict its label

Look up tables achieved in the learning phrase
Decision making
P(Outlook=Overcast |Play=No) = 0
P(Temperature Hot|Play=No) =2/5
P(Humidity=Peak|Play=No) = 4/5
P(Windy=False|Play=No) = 2/5
P(Play=No) = 5/14
P(Outlook=Overcast |Play=Yes) = 4/9
(Temperature=Hot|Play=Yes) = 2/9
P(Humidity= Peak |Play=Yes) = 3/9
P(Windy= False||Play=Yes) = 6/9
P(Play=Yes) = 9/14
P(Yes|x)≈[P(Overcast|Yes)P(Hot|Yes)P(Peak|Yes)P(False|Yes)]P(Play=Yes) = 0.0141093
P(No|x)≈[P(Overcast|No)P(Hot|No)P(Peak |No)P(False|No)]P(Play=No) = 0
Given the fact P(Yes|x) > P(No|x), we label x to be "Play". The new tuple classified under "PLAY" category. Hence there will be play.

III. RESULT AND DISCUSSION

Results are shown in the form of graphical representation. Fig. shows that the accuracy obtained by changing the condition of dataset i. e. S, H, P,F and O, H, P, F contains two distinct values 'yes' and 'no'. The entire content of the probability of play and not play find out from naïve bayes classification method the result of the finding out the probability given in below table. When the probability of play is greater than probability of Not play then it display the output result is "The new tuple classified under " PLAY" category. Hence there will be play." When the probability of no play is greater than Play probability then output result is "The new tuple classified under "NO PLAY" category. Hence there will be NO play."

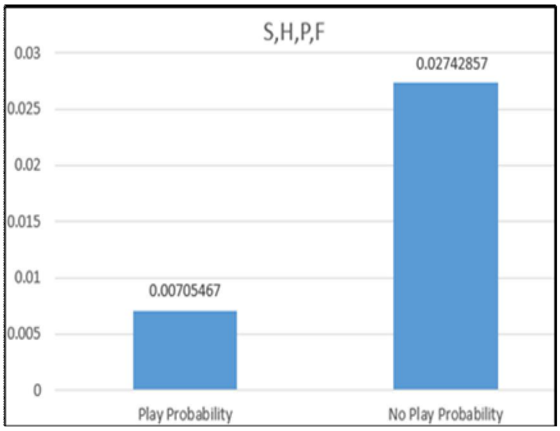

Fig. 2 Result of No play (S,H,P,F)

play 0.00705467 and probability of no play is 0.02742857. Probability of play 0.00705467<probability of no play is 0.02742857. The new tuple classified under "NOT PLAY" category. Hence there will be NO play.

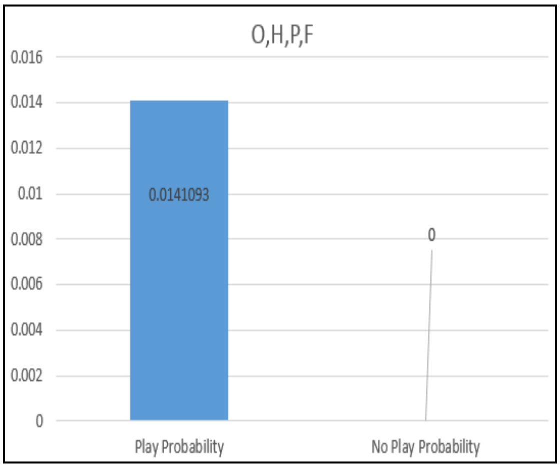

Fig. 3 Result of Play (O,H,P,F)

From above figure 3 it is seen that the probability of play 0.0141093 and probability of no play is0. Probability of play 0.00705467>probability of not play is 0. The new tuple classified under "PLAY" category. Hence there will be play.

Table No. 6 Comparison result of probability

Input	Play Probability (Theorem Calculation)	No Play Probability (Theorem Calculation)	Output (Theorem Calculation)
S,H,P,F	0.00705467	0.02742857	No Play
O,H,P,F	0.0141093	0	Play

IV. CONCLUSION

Use of machine learning getting the hang of utilizing a gullible bayes classifier in examining the play and not play information. It is a decent technique for thinking about the current connections between factors. From our proposed approach we have demonstrated that information mining recovers helpful even from which are not immediate markers of the class we are endeavoring to anticipate. In our work we have attempted to foresee the odds of getting a play and no play likelihood by utilizing credulous bayes classifier. This framework characterizes the given information into various classes and furthermore predicts the danger of the play if obscure Sample is given as an info. The framework can be filled in as device for cricket, Tennis and so football. Additionally, it will help for different games. As we have created summed up framework, in future we can utilize this framework for investigation of various datasets by just changing the name of dataset document which is given for preparing module.

REFERENCES

[1] Mark A. Hall. (2000) 'Correlation-based Feature Selection for Discrete and Numeric Class Machine Learning' in ICML 2000: Proceedings of the Seventeenth International Conference on Machine Learning, Morgan Kaufmann Publishers Inc. San Francisco, CA, USA, pp. 359-366..

[2] Jiawei Han. and Micheline Kambar (2006) Data Mining: Concepts and Techniques, 2nd ed., Morgan Kaufmann Publisher..

[3] Jacek Biesiada and Wlodzislaw Duch (2007) 'Feature Selection for High-Dimesional Data: A Pearson Redundancy Based Filter', Computer Recognition System 2, ASC, Vol. 45, pp.242-249.

[4] SelvaniDeepthiKavilaa, Dr.Radhika Y b, “Research Domain Selection using Naive Bayes Classification”, I.J. Mathematical Sciences and Computing, 2016, 2, 14-23 Published Online April 2016 in MECS (http://www.mecs-press.net) DOI: 10.5815/ijmsc.2016.02.02.

[5] LeniMarlina, Muslim, Andysah Putera UtamaSiahaan, “Data Mining Classification Comparison (Naïve Bayes and C4.5 Algorithms)”, International Journal of Engineering Trends and Technology (IJETT) – Volume 38 Number 7- August 2016, pp. 380-383

[6] K.Mani, P.Kalpana “ An Efficient Feature Selection based on Bayes Theorem, Self Information and Sequential Forward Selection”, I.J. Information Engineering and Electronic Business, 2016, 6, Published Online November 2016 in MECS (http://www.mecs-press.org/)DOI: 10.5815/ijieeb.2016.06.06. pp. 46-54.

[7] Dhanashree S. Medhekar, Mayur P. Bote, Shruti D. Deshmukh, “Heart Disease Prediction System using Naive Bayes”, International journal of enhanced research in science technology & engineering, Vol. 2 issue 3, March.-2013, pp 1-5

[8] Rajeswari R.P, Kavitha Juliet, Dr.Aradhana , “Text Classification for Student Data Set usingNaive Bayes Classifier and KNN Classifier”, International Journal of Computer Trends and Technology (IJCTT) – Volume 43 Number 1 – January 2017 pp. 8-12