IJCSE International Journal of Computer Sciences and Engineering Open Access
Review Paper Vol.-6, Issue-12, Dec 2018 E-ISSN: 2347-2693

Analysis of Naïve Bayes Classification for Diabetes Mellitus

S. Sankaranarayanan[1*], T. Pramananda Perumal[2]

[1]Research and Development Centre Bharathiar University, Coimbatore-46, India
[2](Retd), Presidency College (Autonomous),Chennai-5, India

**Corresponding author: profsankaranarayanan1970@gmail.com, Tel.: +91-7904051822 / 9443651545*

Available online at: www.ijcseonline.org

Accepted: 21/Dec/2018, Published: 31/Dec/2018

Abstract- Data Mining plays a major role in the decision making process of any application as in Health Care, Artificial Intelligence, military and weather forecasting. In particular, Classification is used to implement the real time Clinical Decision Support System (CDSS) in health care industry. Thus the CDSS can be viewed as if it predicts the decisions through the supervised learning instances from the training dataset. Here a discrete set of algorithms and techniques are in vogue in the backdrop of classification through supervised learning and hence termed as classification algorithms. Among these classification algorithms, Naïve Bayes is the most familiar which uses the historical data as supervised learning instances. This paper surveys the application of Naïve Bayes classification in health care with specific pertinence to analyzing Diabetic Mellitus disease. It also focuses on the implementation of this specific algorithm in the Diabetic domain to expertise an application.

Keywords— Classification, Text Classification, Naïve Bayes, Semantic Analysis, Health Care

I. INTRODUCTION

In machine learning, classification can mainly be used for the decision making process in the critical situations and when predictive observations are needed with the application of sample data. Supervised learning instances are used as historical data or training dataset to predict the decisions over the application which does not warrant human intervention. Classification deals with both text-based and semantic content for the decision making analysis. Text classification is obtained for reason prediction through data processing based on Natural Language Processing (NLP). The text classification process consists of five main phases; Data preprocessing, feature extraction, feature selection, classification and performance analysis. Data preprocessing step is performed as being warranted based on the available content and type of the data as to such text or sentence to be made nominal. Here, NLP is used for the preprocessing phase. The whole document or sentence may involve to process of tokenization initially. The tokens can be processed with the procedural tasks such as stop word identification, word sense disambiguation. Finally, stemming process is used to obtain the processed level of data for the input to next phase. Next to preprocessing phase, the feature extraction is used for obtain the key terms for the analysis to predict. The most essential of all the steps is the feature selection phase doing selection of important features/words in the whole document/sentence based on the significance of those keywords in the indispensable classification task. This is done despite the fact that a range of effectual feature selection methods as made available. The initial process is then directing for the final classification phase where the classification algorithms specifically Machine Learning Classifiers are used. The last step finds the effectiveness and performance of the classification task being applied based on the selected classifier and the appropriate feature selection method thus far used.

Section I deals with the introductory concepts on Data mining in the context of classification and is better analyzed both for Text data and Semantic inputs. As machine learning classifiers so selected for classification process dictate the success rate of these algorithms, models implementing various types of classifiers for specific domains are inducted in Section II that follows here. In Section III, special emphasis is given to Naïve Bayesian algorithm as is so important to the context of classification in its simplicity and elegance of exhibiting performance in almost all domains of application including Clinical Decision Support System. It also elucidates all background information of Naïve Bayesian classification thoroughly. Section IV deals with the survey of application of Naïve Bayesian process to various medical sub domains. Since a comparative survey of various classification algorithms for Medical domain is proposed, no numerical inputs have been taken and hence no results are thus far been obtained and discussed.

II. BACKGROUND

There have been many classifiers namely Naïve Bayes[1], Rochhio‘s algorithm[2], Support Vector Machines[3], K-Nearest Neighbor[4], Maximum Entropy classifier[5] etc.

The Naïve Bayes Classifier is a probabilistic graphical model also called as Bayesian Classifier and Bayesian Classification Algorithm which is based on the Bayes Theorem[6] proposed by the Naïve Bayes. Bayesian Classification works on the principle that a class can probably find the features of objects of that class because of the cohesion of this text feature value in the class. If a class is referred to an agent, feature value finding is simpler. Nonetheless, if the class itself is unknown and the only of the content of information are some of the measures of the features, then corresponding class can be predicted by using the Bayes rule and hence the probabilistic model of the obtained features

A learning agent can therefore build a probabilistic model of the available features and use the same for predicting the classification. Naïve Bayes Classification can also be collectively organized in terms of latent variables, i.e. the probabilistic variables are not observed. Thereby building the resultant classification an inference in the probabilistic model in such a case, latent variables are obtained from latent variables observed due to probability. The simplest of all the associated classifiers projected is the Naïve Bayes Classifier. Naïve Bayes classifier fully works on the principle of independent assumption using the input features participating independently. As long as the primary supposition of independence is true, the Naïve Bayes classifier performs as much as good. It is proportional that the obtained class independent of the feature to the assumption for the good feature of the class. When compared to other evolutionary algorithms, it is very simple and easy to implement. The primary hypothesis is not always convenient as per the real time scenario. The Naïve Bayes assumption is also not practical always and also called as linear classifier.

Here it concentrates on the classification of medical data analysis for the prediction of diabetes mellitus from the supervised learning of historical patient data patterns. Diabetes Mellitus is the highly possible as is a predisposed genetic type disease which can be easily predicted through the genetic history. The Clinical Decision Support System is proposed with the data mining techniques such as clustering and classification. As regards to, this survey has been taken for the identification of importance of classification algorithm and a critical comparison of various algorithms including Naïve Bayes Classification algorithm to this medical application are to be made. Major contribution of this survey work is to obtain the preliminaries for the classification and its techniques and prediction of diabetes mellitus. It mainly concentrates on the methodology illustration of the recommended classification algorithm with high accuracy and as is easier to deploy in the medical environment.

III. PRELIMINARIES

A. Classification and Prediction

Data mining is the knowledge discovery technique used to analyze and predict the class labels. It is also called as Knowledge discovery from databases (KDD), is the fully or semi automated pattern extraction strategy from the training data set. Data mining techniques predict the behavior patterns or future trends and used to make the knowledge driven and proactive business decisions for the enhancement of business strategies for higher productivity. The data mining techniques are clustering, classification, generalization and prediction

Classification techniques are mainly used for classifying the data among an assortment of classes in data mining. Classification is widely used to classify the tuple of the industry to identify the type based on the similarities in properties. Classification can be done through the two important processes of Model Construction: Building the classifier model and Model Usage. The constructed classifier model extracts the Class features.

Prediction is the process of obtaining the future trends based on the historical data. It is widely used for forecasting the critical weather situations, predicting the future values in stock trading, identifying the possibilities of medical diseases from the historical patient medical data. Priyanka et al have devised a smart health care system using data mining strategies [7].
In medical health care systems, the diabetes mellitus can be identified and predicted from the patient data set.

B. Naïve Bayes Classification Algorithm

Naïve Bayes Classification is widely applied classification algorithm among all because it supports scalable dimensionality of the input and training data set. Moreover it is simple and more sophisticated amongst classification methods. It displays the probability of each parameter of the input for the conventional state. It implicates from the Bayes Rule to create models with predictive capabilities.

1. Naïve Bayes Rule:

A conditional probability is the likelihood of some conclusion say C, E be some evidence/observation where a dependence relationship exists between C and E.

This probability is denoted as P(C|E) where

$$\mathbf{P(C|E)} = \frac{\mathbf{P(E|C)P(C)}}{\mathbf{P(E)}}$$

2. Naïve Bayesian Classification Algorithm:

The Naïve Bayes Classifier or Linear Bayesian Classifier performs as follows:
Step 1: Let D be tuples collected from training set and their associated class labels Ca and Cp. As usual, each record is represented by an n-dimensional attribute vecor,

X=(x1,x2,…xn-1, xn), representing n measurements made on the tuple from n parameters, A1 to An.

Step 2: Suppose that there are m number of classes for prediction C(C1, C2…, Cm-1,Cm) given a record X, the classifier will predict that X belongs to the class C having the most posterior probability, conditioned on X. The naive bayes rule can be used to predict that the tuple X belongs to the class C

$$P(Ci|X) > P(Cj|X) \quad for\ 1 \le j \le m\ and\ j \ne i$$

Thus we maximize P(Ci|X). The class Ci for which P(Ci|X) when maximum then it is called as maximum posteriori hypothesis by Bayes Theorem

$$P(Ci|X) = \frac{P(X|Ci)P(Ci)}{P(X)}$$

Step 3: As P(X) is constant for all classes, only P(X|Ci) P(Ci) need be maximized. If the class prior probabilities are not known, then it is often assumed that the classes are equally likely, that is P(C1)=P(C2)=…=P(Cm-1) = P(Cm) and it would therefore maximize P(X|Ci). Otherwise, we maximize P(X|Ci)P(Ci). Note that the class prior probabilities may be identified by P(Ci)=|Ci,D|/|D|, where |Ci,D| is the number of training tuples of class Ci in dataset D.

Step 4: Given data sets with many attributes, it would be extremely computationally expensive to compute P(X|Ci). To decrease computation in evaluating P(X|Ci), the naïve assumption of class conditional independence is made. This presumes that the values of the parameters are conditionally independent of one another, given the class label of the tuple. Thus,

$$\mathbf{P(X|Ci) = \prod_{k=1}^{m} P(xk|Ci) = P(X1|Ci) * P(X2|Ci) * \dots * P(xm|Ci)}$$

We can often estimate the probabilities P(X1|Ci),P(X2|Ci)..P(xm|Ci) from the database tuples. For instance, to find P(X|Ci), it is consider that if Ak is categorical, then P(Xk|ci) is the tuples of class Ci in dataset D having the values of Xk for Ak, divided by |Ci,D|, the number of tuples of class Ci in D.

Step 5: In order to find the class label of tuple X, P(X|Ci)P(Ci) is evaluated for each class Ci. The classifier finds that the class label of tuple X is the class Ci if and only if

$$\mathbf{P(X|Ci)P(Ci) > P(X|Cj)P(Cj)\ for\ 1 \le j \le m, j \ne i}$$

The predicted class label for the tuples is the class Ci for which P(X|Ci)P(Ci) is the maximum.

If required for accuracy, the smoothing techniques for Naïve bayes are also used. Smoothing is a technique to make a approximating function used to find the important patterns in the data instead of obtaining the noise or other fine scale structures.

C. Methodology Illustration of Naïve Bayes classifier

Thus the medical disease prediction can be based on the Naïve bayes classification algorithm for the test patient data using the patient medical data history. The illustration of the CDSS using Naïve Bayes Classification is as follows:

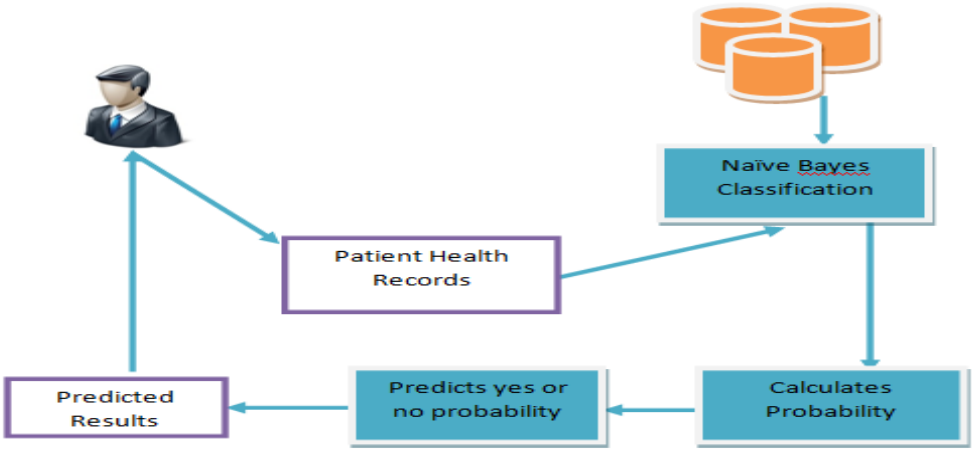

Fig.1. Naive Bayes Classification Methodology

IV. NAÏVE BAYES AND MEDICAL DOMAIN

Borkar and Deshmukh [8] proposed using Naïve Bayes classifier for detection of Swine Flu disease. The process starts with finding probability for each attribute of Swine flu against all output. The probabilities of each attribute are then multiplied. Selecting the maximum probability from all the probabilities, the attributes belong to the class variable with maximum value. The promising results of the proposed scheme can be used for investigating further the Swine flu disease in patients using Information technology.

Patil [9] worked in the direction of diagnosing whether a patient with his given information regarding age, sex, blood pressure, blood sugar, chest pain, ECG reports etc can have a heart disease later in life or not. The experiments involve taking the parameters of the medical tests as inputs. The proposal is effective enough in being used by nurses and

medical students for training purposes. The data mining technique used is Naïve Bayes Classification for the development of Decision Support System in Heart Disease Prediction System (HDPS). The performance of the proposal is further improved using a smoothing operation. The implementation of HDPS is done through a MATLAB application able to detect and extract hidden knowledge related to heart diseases from a historical heart disease database.

Kharya et al [10] proposed detecting in patients the chances of having Breast Cancer later in life. Severity in Breast Cancer is necessary seeing it becoming the second most cause of death among women. A Graphical User Interface (GUI) is designed for entering the patient's record for the prediction. The records are mined through the data repository. Naïve Bayes classifier, being simple and efficient is chosen for the prediction. The results obtained by the Naïve Bayes classifier are accurate, have low computational effort and are fast. Implementation of the proposal is done through Java and the training of data is done using datasets from UCI Machine Repository [11]. Another advantage of the proposed system is that the system expands according to the dataset used.

Stephanie J. Hickey [12] proposed using Naïve Bayes Classifier for public health domain combined with greedy feature selection. The input was a public health dataset and the objective behind the proposal was to identify one or several attributes that best predict a selected target attribute without the need for searching the input space exhaustively. The proposal achieved its goal with increase in accuracy of classification. The target attributes were related to diagnosis or procedure codes.

Ambica et al [13] proposed using Naïve Bayes for an efficient decision support system for Diabetes disease. They proposed classification system was divided into two steps. The first step includes analysis of how optimal the dataset is and accordingly extraction of the optimal feature set from the training data is done. The second step forms the new dataset as the optimal training dataset and the proposed classification scheme is now applied on the optimal feature set. The mismatched and unavailable features from the training and testing datasets are ignored and the dataset attributes are used for the calculation of posterior probability. The proposed procedure therefore shows elimination of unavailable features and document wise filtering.

V. CONCLUSION

Classification algorithms can be selected based on the usage and working principle of specific algorithm. However, the implementation of the algorithm must be as simple as it is with high support for analysis through large number of input parameters. Here, Naïve Bayes Classification algorithm has been explained in detail with its illustrated methodology and algorithmic procedure. The CDSS implementation using classification on various domains of medical data were surveyed. Similar approach is leveled for Medical data mining of Diabetes Mellitus data. In future, the health care system can be implemented for medical data with enhancements through clustering and prediction.

REFERENCES

[1] D. Lewis, ―Naive Bayes at Forty: The Independence Assumption in Information Retrieval‖, Proceedings of the 10th European Conference on Machine Learning(ECML-98), 1998.

[2] Christopher D. Manning, Prabhakar Raghavan and Hinrich Schütze, ―An Introduction to Information Retrieval‖, Cambridge University Press, page 181, 2009.

[3] A. Basu, C. Waters and M.Shepherd, ―Support Vector Machines for Text Categorization‖, Proceedings of the 36th Annual Hawaii International Conference on System Sciences, 2003. For Conference

[4] Gongde Guo, Hui Wang, David Bell, Yaxin Bi and Kieran Greer, ―KNN Model-Based Approach in Classification‖, Proceedings of the ODBASE, pp- 986 – 996, 2003.

[5] Kamal Nigam, John Lafferty and Andrew McCulllum, ―Using Maximum Entropy for Text Classification‖, IJCAI-99, Workshop on Machine learning for Information Filtering, pp. 61-67, 1999.

[6] Stuart, A.; Ord, K. (1994), Kendall's Advanced Theory of Statistics: Volume I—Distribution Theory, Edward Arnold, §8.7.

[7] Priyanka, Sana Khan, Tulsi Kaur—Investigation on Smart Health Care using Data Mining Methods‖, International Journal of Scientific Research in Computer Sciences and Engineering (IJSRCSE), Vol 4, Issue 2, pp 31-36, 2320–088X, 2016

[8] Ankita R. Borkar and Dr. Prashant R. Deshmukh , ―Naïve Bayes Classifier for Prediction of Swine Flu Disease‖, International Journal of Advanced Research in Computer Science and Software Engineering, Volume 5, Issue 4, pp. 120-123, 2277 128X, 2015.

[9] Ms.Rupali R.Patil, ―Heart Disease Prediction System using Naive Bayes and Jelinek-Mercer smoothing‖, International Journal of Advanced Research in Computer and Communication Engineering, Vol. 3, Issue 5, pp. 6787-6792, : 2278-1021 2014.

[10] Shweta Kharya, Shika Agrawal and Sunita Soni, ―Naive Bayes Classifiers: A Probabilistic Detection Model for Breast Cancer‖, International Journal of Computer Applications (0975 – 8887) Volume 92 – No.10, pp.26-31, 2014 [10] UCI Machine Learning Repository, http://ics.uci.edu/ mlearn/MLRepository.html. 0975 – 8887.

[12] Stephanie J. Hickey, ―Naive Bayes Classification of Public Health Data with Greedy Feature Selection‖, Communications of the IIMA, Volume 13, Issue 2 Article 7, pp. 87-98, 2013.

[13] A.Ambica, Satyanarayana Gandi and Amarendra Kothalanka, ―An Efficient Expert System For Diabetes By Naïve Bayesian Classifier‖, International Journal of Engineering Trends and Technology (IJETT) –Volume 4 Issue 10, pp.4634-4639, 2231-5381 2013

[14] Pablo Gamallo, Marcos Garcia and Santiago Fernández-Lanza, ―TASS: A Naive-Bayes strategy for sentiment analysis on Spanish tweets‖, Workshop on Sentiment Analysis at SEPLN (TASS2013), pp. 126-132, 2013.

Authors Profile

Mr. S.Sankaranarayanan had his Master degree in Computer Science from Bharathidasan University, Tiruchirappalli in 1991and Master of Philosophy in Computer Science from Manonmaniam Sundaranar University, Tirunelveli in the year 2003. He is currently pursuing his Ph.D in Bharathiar University, Coimbatore and currently working as Associate Professor in the Department of Computer Science, Government Arts College, Kumbakonam since 1999. He has published more than 10 research papers in reputed international journals and conferences including IEEE and it's also available online. His main research work focusses on Medical Data Mining, Natural Lanaguage Processing and Cloud Computing systems. He has more than 20 years of teaching experience and 5 years of Research Experience.

Mr T.Pramanandaperumal has got graduated to Master of Science in Physics from Madurai Kamaraj University, India in the year 1981 and had his Master of Philosophy in Computer Science from University of Madras in 1986. He pursued his Ph.D with University of Madras in the year 2008 and was working as Associate Professor and Head in the Department of Computer Science in various Government Colleges of Tamilnadu and also served as Principal in Four Government Colleges of Tamilnadu including the pretegious Presidency College of Chennai. He was a Member of Broard of Studies in various Universities and Colleges and during his period of service as Chairman of Board of Studies in Computer Science in University of Madras once. He has published more than 20 research papers in reputed international journals including Elsevier and conferences including IEEE and it's also available online. His main research work focusses on Computer Simulation modelling, Computer Graphics, Image Processing and Data Mining. He has more than 35 years of teaching experience and 18 years of Research Experience.

