An Efficient Expert System For Diabetes By Naïve Bayesian Classifier

A.Ambica[1], Satyanarayana Gandi[2] , Amarendra Kothalanka[3]
M.Tech scholar[1], Associate Professor[2], Professor[3]
[1,3]Department of Computer Science & Engineering,[2]Department of Information Technology,
[1,2,3]Dadi Institute of Engineering and Technology (Affiliated to JNTUK), Anakapalle, Andhra Pradesh

Abstract:- In this paper we are proposing an efficient decision support system for Diabetes Disease, apart from the traditional simple support vector machine. We are proposing an efficient two level approach for classifying data. In initial phase we extract optimal feature set from the training data by analyzing the optimality in the dataset, then new dataset is formed as optimal training dataset, now we apply our classification mechanism on the optimal feature set.

I.INTRODUCTION

Researchers developed a fuzzy-based controller that incorporates expert knowledge to regulate the blood glucose level. Magni and Bellazzi [2] devised a stochastic model to extract variability from a self monitoring blood sugar level time series. Polat and Gunes [3] designed an expert system to diagnose the diabetes disease based on principal component analysis. Polat et al. [4] also developed a cascade learning system to diagnose the diabetes. Chang and Lilly [5] developed an evolutionary approach to derive a compact fuzzy classification system. Goncalves et al. [6] introduced an inverted hierarchical neuro-fuzzy BSP system for pattern classification and rule Extraction in databases and Kahramanli and Allahverdi [7] designed a hybrid neural network system for classification of the diabetes database. Chang-Shing Lee [8] designed as fuzzy expert system for diabetes decision support application based on the fuzzy ontology with five layer fuzzy ontology. Ismail saritas et al.[9] developed a fuzzy expert system to determine drug dose in treatment of chronic intestine inflammation using the concept of fuzzification. Mehdi Fasanghari et al.[10] developed a fuzzy expert system for Tehran stock exchange using the concept of fuzzification. Diabetes treatment focuses on controlling blood sugar levels to prevent various symptoms and complications through diet and exercise. The American Diabetes Association [11] categorizes diabetes into type-1 diabetes, which is normally diagnosed in children and young adults, and type-2 diabetes, i.e., the most common form of diabetes that originates from a progressive insulin secretary defect so that the body does not produce adequate insulin or the insulin does not affect the cells. The Bayesian classification easing number of diabetics worldwide has drawn the attention of a diverse array of fields, including artificial intelligence and biomedical engineering, explaining why related technologies such as fuzzy inference mechanisms and fuzzy expert systems have been adopted for diabetes research.

More number of the studies have shown that patients suffering from Diabetes can significantly delay the onset and slow down the progression of diabetes micro- and macro-angiopathic complications through intensive treatment and monitoring as In general intensive treatments imply a careful blood case of glucose level (BGL) self-monitoring process of analysis of BGL measurements is one of the most important tasks in order to assess the glucose metabolic control and to revise the therapeutic model in Recent clinical studies have shown the correlation between the glucose variability and the long-term diabetes related complications.

II. RELATED WORK

Expert knowledge system has interesting research work during these years, specifically in the medical field of Diabetes mellitus. This illness requires continuous and regular treatment for the patient who are suffering with Diabetes mellitus, researchers in a way to find an optimal solution of expert knowledge system .Most of the Traditional knowledge systems and classifications works with probability densities density variation between the training and testing datasets .These mechanisms suffering with so many drawbacks like new attribute recovery and mismatch of the the attribute in the training data set and testing dataset , because in there time environment end user analyst cannot expect the semantic so the attributes of the training and testing datasets So, if analyst can reduce the computational complexity regarding mismatched attributes and We are rectifying one more drawback in the classification approach with elimination features, we will ignore the mismatched unavailable features from the training and testing datasets and calculates the posterior probability with the attributes of the datasets, the following procedure shows the document wise filtering and elimination. We integrated optimal extraction for the optimal diabetes results by eliminating the unavailable attributes from the training datasets and testing datasets

III. PROPOSED SYSTEM

In this proposed approach we are introducing two level approaches. Initially we Extract the optimal feature set from the existing training data and calculate the positive and negative probability, until a new dataset is formed with same size and forwards the current generated dataset for the classification; there it classifies the testing data features with the new Dataset.

A) Optimal Feature Extraction

Traditional approaches of knowledge expert system works with static measures and it may contains unnecessary information, which means the attributes doesn't satisfies the minimum threshold value. In our proposed approach Dataset is gathered for the decision support system, with relevant diabetes characteristics or feature sets and before forwarding the dataset to the classification ,forward the Dataset to the Optimal feature set selection process with the specified threshold values of the Dataset.

The objects satisfies the minimum threshold value can be treated as positive attributes and other can be treated as negative attributes, for Optimal extraction of the datasets, remove the records which contains the negative attributes, forward the remaining dataset as optimal dataset to further classification process.

B) Baye's Theorem

Bayes theorem is a simple calculation of finding probability factors over existing and new attribute possible values over the samples .This approach shows the how the probability works with theoretical values of the samples, Bayesian can be applied in many ways while there is a possibility of existing attribute values and new samples and can be measures in terms of probability class labels.There are so many real time application which are using bayseian approach bank relation, corporate analyzes, educational analysis, science and technology etc..

Notation of the bayes approach is given as follow with sample notation of the probability and with its conditional probability with respect to event

$$P(T|E)=\frac{P(E|T)\times P(T)}{P(E|T)\times P(T)+P(E|\sim T)\times P(\sim T)}$$

In this below formula, T indicates the theory or hypothesis that we are interested in testing E represents a new piece of evidence that seems to confirm or disconfirm the theory. For any proposition S it will use P(S) to stand for our degree of belief, or "subjective probability," that S is true. Particularly P(T) represents our best estimate of the probability of the theory we are considering and prior to consideration of the evidence. It is known as the *prior probability* of T. Our work is to discover is the probability that T is true supposing that our new piece of evidence is true. This is a *conditional probability* and the probability is that one proposition is true provided that another proposition is true. Suppose you draw a card from a deck of 52 and without showing it to me. Consider that the deck has been well shuffled and I should believe that the probability that the card is a jack and P(J) is 4/52 or 1/13 since there are four jacks in the deck. Suppose you tell me that the card is a face card. The resultant probability that the card is a jack and it is given that it is a face card and 4/12, or 1/3, since there are 12 face cards in the deck. We denote this conditional probability as P(J|F), meaning the probability that the card is a jack *given that* it is a face card. (It don't need to take conditional probability as a primitive notion; we define it in terms of absolute probabilities: P(A|B) = P(A and B) / P(B) that is the probability that A and B are both true divided by the probability that B is true.)

Using this idea of conditional probability to express what we want to use Bayes' Theorem to discover, we say that P(T|E), the probability that T is true given that E is true and it is the *posterior probability* of T. The idea is that P(T|E) represents the probability assigned to T *after* taking into account the new piece of evidence E. To find the value we need addition to the prior probability P(T) and two further conditional probabilities indicating how probable our piece of evidence is depending on whether our theory is or is not true. We can represent these as P(E|T) and P(E|~T) where ~T is the *negation* of T i.e. the proposition that T is false. Proposed approach as follows

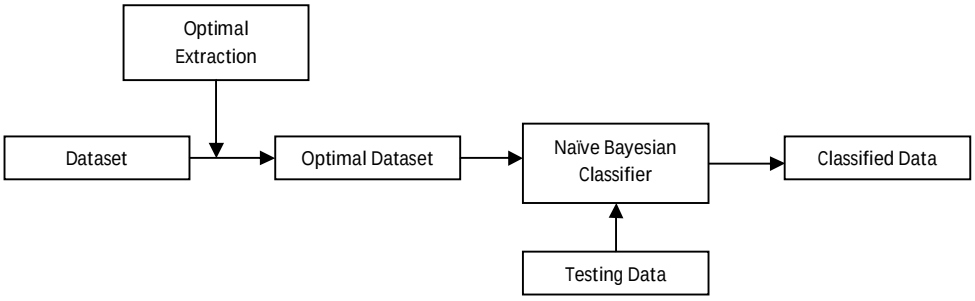


Figure1: Proposed architecture

C) Classification

For the Classification mechanism, we are using Bayesian classifier for classify the testing dataset with newly formed optimal feature set based Dataset for diabetes. This approach works with corresponding posterior probability of the individual features with respect to the original dataset.

For the classification process we are using Bayesian classifier for analyzing the testing data with the training information. Bayesian classifier is defined by a set C of classes and a set A of attributes. A generic class from C is denoted by c_j and a generic attribute belonging to A as A_i.Consider a database D with a set of attribute values and the class label of the case. The training of the Naïve Bayesian Classifier consists of the estimation of the conditional probability distribution of each attribute is given in the class.

Here P(X) is prior probability =

P is the data sample from our set of fruits is red and round) P(X), P(H), and P(X/H) may be estimated from given data .Use of Bayes Theorem in Naïve Bayesian Classifier

1. Each data sample is of the type

   $X=(x_i)$ i =1(1)n, where x_i is the values of X for attribute A_i

2. Consider we have m classes C_i, i=1(1)m.

   $X \in C_i$ iff

   $P(C_i|X) > P(C_j|X)$ for $1 \leq j \leq m, j \neq i$

   i.e BC assigns X to class C_i having highest posterior probability conditioned on X . The class for which $P(C_i|X)$ is maximized is called the maximum posterior hypothesis. From Bayes Theorem

3. P(X) is constant. Only need be maximized.

   - If class prior probabilities not known and assume all classes to be equally likely
   - else maximize

$P(C_i) = S_i/S$

Problem: computing $P(X|C_i)$ is unfeasible!

(find that how would you find it and why it is infeasible)

4. Naïve consideration is: attribute independence

   $= P(x_1,\ldots,x_n|C) = \Pi P(x_k|C)$

5. In order to classify an unknown sample X, evaluate for each class C_i. Sample X is included to the class C_i iff

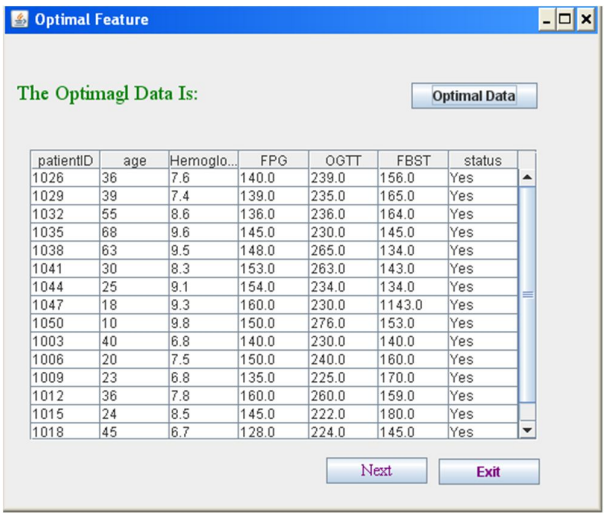
$P(X|C_i)P(C_i) > P(X|C_j)\ P(C_j)$ for $1 \leq j \leq m, j \neq i$.

Experimental Analysis:

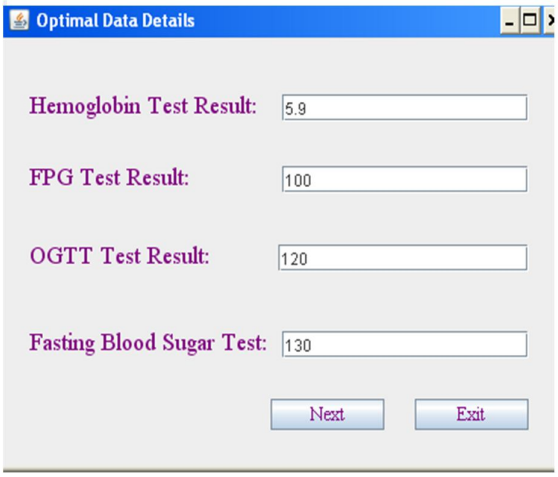
Our implementation purpose we have used language java and some synthetic datasets for analysis, the following representation shows the complete implementation of the architecture.

patientID	age	Hemoglo...	FPG	OGTT	FBST	status
1025	27	6.2	105.0	168.0	118.0	Pre
1026	36	7.6	140.0	239.0	156.0	Yes
1027	23	5.2	92.0	129.0	78.0	No
1028	33	6.3	107.0	168.0	104.0	Pre
1029	39	7.4	139.0	235.0	165.0	Yes
1030	28	5.2	96.0	135.0	84.0	No
1031	45	5.7	109.0	178.0	102.0	Pre
1032	55	8.6	136.0	236.0	164.0	Yes
1033	60	5.6	76.0	104.0	60.0	No
1034	65	5.9	104.0	100.0	115.0	Pre
1035	68	9.6	145.0	230.0	145.0	Yes
1036	70	5.3	73.0	120.0	65.0	No
1037	53	6.3	123.0	163.0	112.0	Pre
1038	63	9.5	148.0	265.0	134.0	Yes
1039	62	4.9	88.0	106.0	70.0	No
1040	50	5.9	113.0	178.0	123.0	Pre
1041	30	8.3	153.0	263.0	143.0	Yes
1042	20	5.2	67.0	126.0	87.0	No
1043	22	5.9	104.0	183.0	122.0	Pre

The above figure shows the Synthatic dataset before optimal extraction process and it it will be forwarded to optimal extraction process as follows

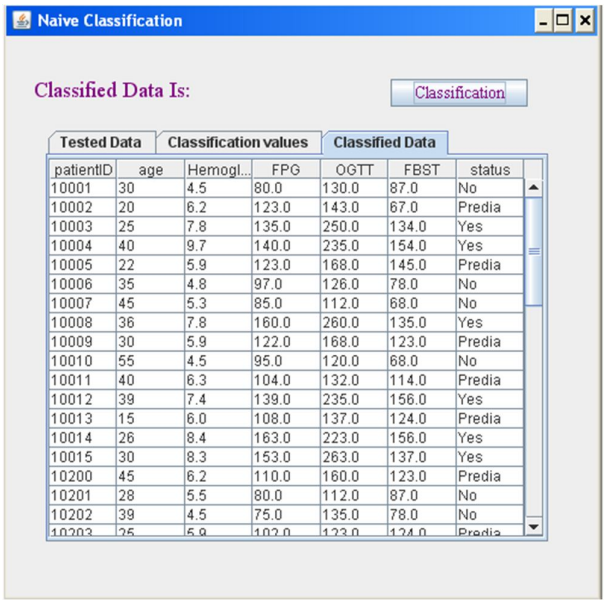


patientID	age	Hemoglo...	FPG	OGTT	FBST	status
1026	36	7.6	140.0	239.0	156.0	Yes
1029	39	7.4	139.0	235.0	165.0	Yes
1032	55	8.6	136.0	236.0	164.0	Yes
1035	68	9.6	145.0	230.0	145.0	Yes
1038	63	9.5	148.0	265.0	134.0	Yes
1041	30	8.3	153.0	263.0	143.0	Yes
1044	25	9.1	154.0	234.0	134.0	Yes
1047	18	9.3	160.0	230.0	1143.0	Yes
1050	10	9.8	150.0	276.0	153.0	Yes
1003	40	6.8	140.0	230.0	140.0	Yes
1006	20	7.5	150.0	240.0	160.0	Yes
1009	23	6.8	135.0	225.0	170.0	Yes
1012	36	7.8	160.0	260.0	159.0	Yes
1015	24	8.5	145.0	222.0	180.0	Yes
1018	45	6.7	128.0	224.0	145.0	Yes



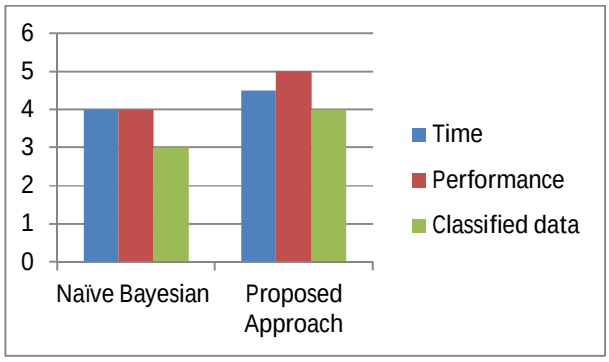
In the above screen we will specify the threshold values of the attributes,then optimal extraction can be done(i.e retrieves the record which satisfies the threshold values).

Now we will forward the optimal dataset to our naïve classification approach for classifying the testing data with training data.



patientID	age	Hemogl...	FPG	OGTT	FBST	status
10001	30	4.5	80.0	130.0	87.0	No
10002	20	6.2	123.0	143.0	67.0	Predia
10003	25	7.8	135.0	250.0	134.0	Yes
10004	40	9.7	140.0	235.0	154.0	Yes
10005	22	5.9	123.0	168.0	145.0	Predia
10006	35	4.8	97.0	126.0	78.0	No
10007	45	5.3	85.0	112.0	68.0	No
10008	36	7.8	160.0	260.0	135.0	Yes
10009	30	5.9	122.0	168.0	123.0	Predia
10010	55	4.5	95.0	120.0	68.0	No
10011	40	6.3	104.0	132.0	114.0	Predia
10012	39	7.4	139.0	235.0	156.0	Yes
10013	15	6.0	108.0	137.0	124.0	Predia
10014	26	8.4	163.0	223.0	156.0	Yes
10015	30	8.3	153.0	263.0	137.0	Yes
10200	45	6.2	110.0	160.0	123.0	Predia
10201	28	5.5	80.0	112.0	87.0	No
10202	39	4.5	75.0	135.0	78.0	No
10203	25	5.9	102.0	123.0	124.0	Predia

Performance representation of the traditional naïve Bayesian and our optimal approach shown as below



Performance analysis

IV.CONCLUSION

We conclude that the project our system provides an efficient knowledge expert system by the naïve classification, in our proposed approach instead of classifying the traditional testing data with training data, we forward the initial training data to the optimal process, to extract the optimal data set, on that optimal dataset we apply classification with Bayesian classifier.

REFERENCES

[1] D. U. Campos-Delgado, M. Hernandez-Ordonez, R. Femat, and A. Gordillo-Moscoso, .Fuzzy-based controller for glucose regulation in type-1 diabetic patients by subcutaneous route,.*IEEE Trans. Biomed. Eng.*, vol.53, no.11, pp.2201.2210, Nov. 2006.

[2] P. Magni and R. Bellazzi, .A stochastic model to assess the variability of blood glucose time series in diabetic patients self-monitoring,. *IEEE Trans. Biomed. Eng.*, vol. 53, no. 6, pp. 977.985, Jun. 2006.

[3] K. Polat and S. Gunes, .An expert system approach based on principal component analysis and adaptive neuro-fuzzy inference system to diagnosis of diabetes disease,. *Dig. Signal Process.*, vol. 17, no. 4, pp. 702.
710, Jul. 2007.

[4] E. D. Lehmann, "Application of computers in clinical diabetes care,"*Diab. Nutr. Metab*, vol. 10, pp. 45–59, 1997.

[5] M. G. Kahn, C. A. Abrams, and M. J. Orland, "Intelligent computerbased interpretation and graphical presentation of self-monitored blood glucose and insulin data," *Diab. Nutr. Metab.*, vol. 4, pp. 99–107, 1991.

[6] S. Andreassen, J. Benn, R. Hovorka, K. G. Olesen, and E. R. Carson, "A probabilistic approach to glucose prediction and insulin dose adjustment: description of metabolic model and pilot evaluation study,"
Comput. Meth. Programs Biomed., vol. 41, pp. 153–165, 1994.

[7] R. Bellazzi, P. Magni, and G. De Nicolao, "Bayesian analysis of blood glucose time series from diabetes home monitoring," *IEEE Trans.Biomed. Eng.*, vol. 47, no. 7, pp. 971–975, Jul. 2000.

[8] S. Montani, P. Magni, R. Bellazzi, C. Larizza, A. V. Roudsari, and E. R. Carson, "Integrating model-based decision support in a multi modal reasoning system for managing type 1 diabetic patients," *Arti. Intell. Med.*, vol. 29, pp. 131–151, 2003.

[9] J. S. Naylor, A. S. Hodel, A. M. Albisser, J. H. Evers, J. H. Strickland, and D. A. Schumacher, "Comparison of parametrized models for computer-based estimation of diabetic patient glucose response," *Med. Inform.*, vol. 22, no. 1, pp. 21–34, 1997.

[10] R. Bellazzi, C. Larizza, P. Magni, S. Montani, and M. Stefanelli, "Intelligent analysis of clinical time series: an application in the diabetes mellitus domain," *Artif. Intell. Med.*, vol. 20, no. 1, pp. 37–57, 2000.

[11] R. Bellazzi, M. Arcelloni, P. Ferrari, P. Decata, M. E. Hernando, A. Garcia, C. Gazzaruso, E. J. Gomez, C. L. C. P. Fratino, and M. Stefanelli, "Management of patients with diabetes through information technology: tools for monitoring and control of the patients' metabolic behavior," *Diab. Technol. Ther.*, vol. 6, pp. 567–578, 2004.

[12] C. L. Rohlfing, H. Wiedmeyer, R. R. Little, J. D. England, A. Tennill,
and D. E. Goldstein, "Defining the relationship between plasma glucose
and HbA1c: analysis of gluose profiles and HbA1c in the diabetes control and complications trial," *Diab. Care*, vol. 25, pp. 275–278,
2002.

[13]http://www.ic.unicamp.br/~rocha/teaching/2011s2/mc906/aulas/naive-bayes-classifier.pdf

[14] B. P. Kovatchev, D. J. Cox, A. Kumar, L. Gonder-Frederick, and W. Clarke, "Algorithmic evaluation of metabolic control and risk of severe hypoglycemia in type 1 and type 2 diabetes using self-monitoring blood glucose data," *Diab. Technol. Ther.*, vol. 5, pp. 817–828, 2003.

[15] M. Muggeo, G. Verlato, E. Bonora, G. Zoppini, M. Corbellini, and E. de Marco, "Long-term instability of fasting plasma glucose, a novel predictor of cardiovascular mortality in elderly patients with noninsulin-
dependent Diabetes Mellitus. The Verona diabetes study," *Circulation*, vol. 96, pp. 1750–1754, 1997.

[16] M. Muggeo, G. Zoppini, E. Bonora, E. Brun, R. Bonadonna, P. Moghetti, and G. Verlato, "Fasting plasma glucose variability predicts 10-year survival of type 2

diabetic patients: the Verona diabetes study," *Diabetes*, vol. 23, pp. 45–50, 2000.

[17] I. B. Hirsch and M. Brownlee, "Should minimal blood glucose variability become the goal standard of glycemic control?," *J. Diab. Its Complications*, vol. 19, pp. 178–181, 2005.

[18] F. J. Service, G. D. Molnar, J. W. Rosevear, E. Ackerman, L. C. Gatewood,
and W. F. Taylor, "Mean amplitude of glycemic excursions, a measure of diabetic instability," *Diabetes*, vol. 19, pp. 644–655, 1970

BIOGRAPHIES

A.Ambica completed her MSC (Computer Science), and currently she is currently pursuing M.Tech in Department of CSE in Dadi institute of Engineering and Technology. Her interested areas are data mining and data warehousing.

Satyanarayana Gandi is an Associate Professor & Head of the Department of Information Technology, Dadi Institute of Engineering and Technology (Affiliated to JNTUK), Anakapalle, Andhra Pradesh, India. He obtained his M.Tech. in Computer Science & Engineering from Andhra University. He is pursuing Ph.D. in Computer Science & Engineering from GITAM University, Visakhapatnam. His main research interests are Safety critical systems, computer Networks, Service computing. He is the Student Branch Counselor of the DIET CSI Student Branch.

Amarendra Kothalanka is a Professor & Head of the Department of Computer Science & Engineering, Dadi Institute of Engineering and Technology (Affiliated to JNTUK), Anakapalle, Andhra Pradesh, India. He obtained his M.Tech. in Computer Science & Technology from Andhra University. He is pursuing his Ph.D in Computer Science & Engineering from GITAM University,Visakhapatnam. His main research interests are Safety Critical Computer Systems, Software Engineering and Mobile Computing. He is the Sponsor of DIET ACM Student, Women in Computing and Professional Chapters.