International Journal on Advanced Science Engineering Information Technology

Vol.10 (2020) No. 5
ISSN: 2088-5334

A Novel Fuzzy Linguistic Fusion Approach to Naive Bayes Classifier for Decision Making Applications

Bahari T. Femina[a,1], Elayidom M. Sudheep[a,2]

[a] Department of Computer Science & Engineering, Cochin University of Science & Technology, Kochi, Kerala-682022, India.
E-mail:[1]feminabahari@gmail.com; [2]sudheepelayidom@gmail.com

Abstract— **Naive Bayes is one of the most widely used classifier algorithms in various data mining problems. The performance of the Naïve Bayes Classifier is comparable to other classifiers as it yields impressive results in multiple applications. An increase in the performance of the Naive Bayes Classifier is possible by identifying and forming segments of the data handled by the classifier. In this paper, a novel fuzzy-based fusion approach to selected quantitative features is proposed. The approach is used to improve the prediction accuracy of the Naive Bayes Classifier (NBC). The linguistic computing model with fusion operators, using ranked indexes of the linguistic terms in the dataset is made use in this proposed approach. Fuzzy values are generated only for the numerical attributes in the initial phase using 2-tuple linguistic computations. The equivalent real value computations are performed in order to express the results in the initial domain of the expression. These computations ensure improved comprehensiveness of the results of the classifier. The model incorporates the concepts of linguistic terms, fuzzy logic, fusion methods, and aggregation operations to the classical Naïve Bayes Classifier. Such incorporation is used to improve the performance of the classifier in various decision-making applications. The proposed model is validated using a standard benchmark dataset–Stat log Heart disease dataset. It is obtained from the UCI Machine Learning Repository. The proposed Linguistic Fuzzy Naive Bayes Classifier showed better accuracy compared to the Simple Naive Bayes Classifier performance.**

Keywords—**naive bayes; fuzzy; linguistic; uncertainty.**

I. Introduction

Naive Bayes is suitable for many decision-making applications in the field of bioinformatics, medicine, business, education, text classification, supplier segmentation, pattern recognition. In Naive Bayes, along with the probabilistic approach, the assumption of independence among the attributes makes it more simple, effective, and robust [1]. The rules or methods devised for prediction in many decision-making applications can be made more useful and accurate by incorporating human knowledge into it. There are various techniques in a classification that allow the impact of human understanding to be found out in the decisions made. Naive Bayes and Bayes Network methods work well with this approach [2].

As the fuzzy set theory is used, the approach is "fuzzy." The terms such as "good," "very good," and "poor" and so on are used to describe the vagueness and uncertainties in the decision maker's thoughts. Hence the approach is termed as "linguistic." Finally, the use of fusion operators in the computational steps explains the use of the term "fusion." The linguistic description is transformed into a linguistic computational model with the definition of membership functions. Zadeh uses the fuzzy set theory to deal with human uncertainties and vagueness in concepts, methods, and decisions [3]. The linguistic representation of variables, thus generating linguistic terms related to applications, creates a method to improve human comprehensibility [4]. In many applications, it was noted that the precise numerical values alone could not be considered as a means of accurate assessment.

When qualitative aspects can be added to a certain phenomenon, the assessment can be made more effective. The Computing with Word (CWW) approach was used in many such decision-making applications by introducing different granularities of uncertainties [4]-[7]. Reasoning by human beings use local information rather than global information. This kind of approach gives rise to some degree of consistency. The use of consistency indexes in decision-making applications is adopted in many areas [8]-[11].

The use of aggregation operators based on priority and generalization of the mean for both triangular and trapezoidal fuzzy information can be seen in many models [12]-[14]. In the intuitionistic fuzzy numbers approach, the numbers are used to represent all the pairwise comparison judgment information over the objects [15], [16]. Multiattribute group decision-making methods also use various aggregation operators, intuitionistic fuzzy values, and intuitionistic 2-tuple linguistic information [17]–[21].

In various application areas such as information retrieval and accessing systems, supply chain risk analysis, engineering systems, credit analysis, and medical diagnosis, the computing with word approach is used to deal with vagueness and uncertainty issues [22]–[25]. A conservative fuzzy logic extension of Naive Bayes Classifier used for incremental learning was proposed by Storrs [26]. It was fast, capable of dealing with missing attributes and the approach behaves exactly as a Naive Bayesian Classifier when the membership function assumes values in [0, 1]. Xi proposed a Fuzzy Naive Bayesian classifier with weights and without restriction for regulated relations [27]. Various versions of the Fuzzy Naive Bayes method using rules and member functions are used in different applications. Tang et al. proposed a Fuzzy Naive Bayes method with a fuzzy clustering algorithm that determines partitions in the space of decision, and these partitions were used as parameters for linguistic variables [28]. This method reduces the learning complexity of the Naive Bayes Method and makes possible the use of continuous variables. In another approach, a method to identify a fuzzy model from data is presented by using the Fuzzy Naive Bayes and a real-valued genetic algorithm. The real-valued genetic algorithm is incorporated to improve the accuracy of the model. The membership functions occurring in the rules are optimized in this model [29]. In another approach, an Aggregated Fuzzy Naive Bayes Data Classifier was proposed as an improved version of the Fuzzy Naive Bayes Classifier and simple NBC. The theoretical part of the proposed classifier in this method is based on arithmetic operations using Chen's Function principle [25].

Doctors make use of various signs and symptoms and other tests for the diagnosis of heart problems in patients. An expert doctor can always provide a better insight into the critical factors that contribute to heart disease prediction. With the help of relevant data and the associated studies conducted on it, prediction can be made on a newly admitted patient. With the help of expert doctors, one can define member functions for the medical factors relevant to heart diseases. Many papers use various data mining techniques for predicting heart diseases.

Many papers conducted studies on Naive Bayes and its variations for improving performance using the Stat log Heart Dataset in UCI Machine Learning Repository [UCI]. A prototype Intelligent Heart Disease Prediction System (IHDPS) was developed using data mining techniques, namely, Decision Trees, Naïve Bayes, and Neural Network. Analysis of results shows that each technique is unique and has enough strength to achieve the objectives of the defined mining goals [30].

An efficient approach for the extraction of significant patterns using the MAFIA algorithm was proposed by Patil. Here K-means clustering algorithm is used to extract the data relevant to a heart attack before frequent pattern mining [31]. Yet another probabilistic approach with Naive Bayes and improvement with supervised equal frequency discretization of numerical data is proposed by Bonaick [32]. Many other techniques using neural networks, fuzzy logic, and genetic engineering are also used in the prediction of heart failure disease [33]–[35].

In this paper, a novel Fuzzy Linguistic Fusion approach to Naive Bayes Classifier in decision-making applications is proposed. The selected numerical attributes are fuzzified with associated multi granular linguistic information, defined member functions, and the fusion operators. The fuzzified values are expressed in the real value domain to incorporate various classifier implementation techniques. Stat log Heart Dataset from UCI Machine Learning Repository is used for the experimental analysis. Python and WEKA tool is used for the computation and prediction of the experimented data. The proposed approach performs computations in two phases. The basic concepts of Linguistic Fuzzy Set, membership functions, and linguistic descriptors are discussed in section I. Section II describes the Linguistic Computational Models, the definitions of the basic operations and functions used in the Fusion approach and the Fuzzy Naive Bayes Classifier. The computational framework of the proposed Linguistic Fuzzy Naïve Bayes Classifier (LFNBC) is explained in section IV. Section V describes the experimental setup, including dataset description, implementation logic and the analysis of the results. Section VI covers the conclusion and the future scope of the work done.

II. Materials and method

The basic concepts and definitions used in the proposed framework and approach are explained in this section. This section also covers the different steps involved in the proposed method adopted for the study of the Fuzzy Linguistic Naive Bayes Classifier. The concepts of fuzzy linguistic set, the different linguistic computational models, definitions, and Fuzzy Naive Bayes Classifiers are explained before proposing the computational framework.

A. Linguistic Fuzzy Set

Linguistic variables are words described in natural language. These variables serve the purpose of describing a concept that cannot be fully defined in quantitative terms. The vague thoughts and decisions are represented using the fuzzy theory [3]. A pair *(F, μ)* defines a fuzzy set where F is a set and *μ: F→ [0, 1]* is a function. For each $x \in F$, *μ(x)* is called a membership function of *x* in *(F, μ)*. A triangular membership function defined with parameters (a, b, c) is as shown in Fig.1.

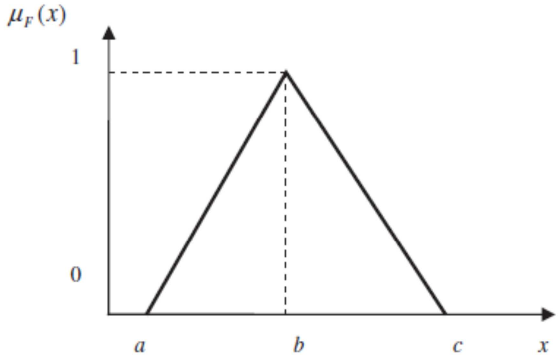

Fig. 1. Triangular membership function μ with parameters (a, b, c)

The triangular fuzzy membership function μ$_F$ (x) is defined as in the following equation (1) and the parameters are defined by a 3-tuple (a,b,c).

$$\mu_F(x) = \begin{cases} 0, & x \leq a \\ \dfrac{x-a}{b-a}, & a \leq x \leq b \\ \dfrac{c-x}{c-b}, & b \leq x \leq c \\ 0, & c \leq x \end{cases} \qquad (1)$$

Fuzzy linguistic model is a good choice in decision-making applications where the quantitative values can be represented and manipulated by computations involving qualitative concepts. The approach of using membership function along with linguistic variables, is used in many applications. This approach has the advantage of approximately characterizing the concept in place of a crisp definition. The concept of vagueness and uncertainties are explained using linguistic terms such as “very low,” “medium,” “and very high” and so on. These linguistic descriptors are human-understandable and easily interpretable. Typically, linguistic descriptors are odd number terms ranging from values 3, 5, 7, 9 and so on. For a set of 7 terms, the midterm value is 0.5, and other values are placed symmetrically around it. Such a set can be defined, as shown in (2). *LT* denotes the linguistic term set.

$$LT = \{LT_0 = Nothing, LT_1 = Very\ Low, LT_2 = Low, LT_3 = Medium, LT_4 = High, LT_5 = Very\ High, LT_6 = Perfect\} \qquad (2)$$

Let four different alternatives of an attribute be represented by $X = \{x1, x2, x3, x4\}$ in which each of the value is defined using the linguistic term set as defined in 2. A linguistic term set of 7 labels is represented in Fig. 2. The representations of these alternatives using different computational models are mentioned in the next section.

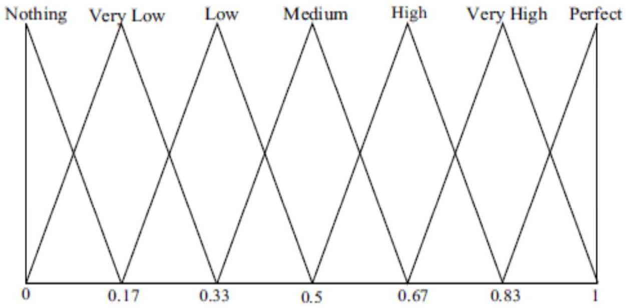

Fig.2. Linguistic term set of 7 labels.

B. Linguistic Computational Models

There are different approaches in the literature to compute the linguistic information [5], [6]. The semantic model is a model in which the computation of fuzzy numbers is based on the fuzzy extension principle. In a symbolic model, the computations are based on the index of the labels. The third model is an extension to the second model, which is a 2-tuple linguistic fuzzy model [36].

1) Semantic Model: The first computational model is based on the semantics and the defined membership functions of linguistic terms. The use of the extension principle in the computation increases the vagueness of the results. The approximation function leads to a lack of accuracy in results. The output can be either fuzzy numbers or linguistic labels. The results of such computations are as expressed in Fig. 3.

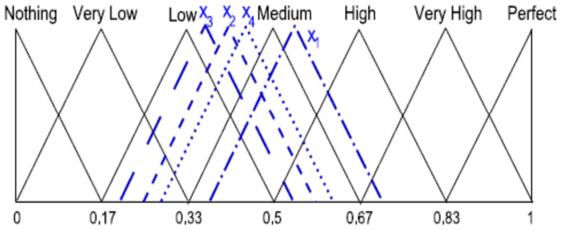

Fig. 3. Results from computational models based on the membership function.

2) Symbolic Model: The second computational model that involves symbolic computations uses the ordered structure of the linguistic term set. Computations are performed on the labels. The classical operators such as Max, Min and Neg are used for aggregating the information. Another variation of the same model uses the convex combination of labels. This model assumes odd cardinality of linguistic terms and the labels are arranged symmetrically on either side of the middle term in the term set. This model is like the representation in Fig. 2. Here the results are also fuzzy numbers. In all these models, the results after computations do not match with the labels in the initial linguistic term set. Hence an approximation process is required. This approach results in a loss of information to a large extend [37].

3) Tuple Linguistic Computational Model: In order to avoid the computational limitations in semantic and symbolic models, the 2-tuple fuzzy linguistic representation model was introduced with extended symbolic computations [36]. The model extends the use of the index of the labels. The accuracy is improved by the addition of a parameter to the basic linguistic representation. The results obtained in this model are as shown in Fig. 4.

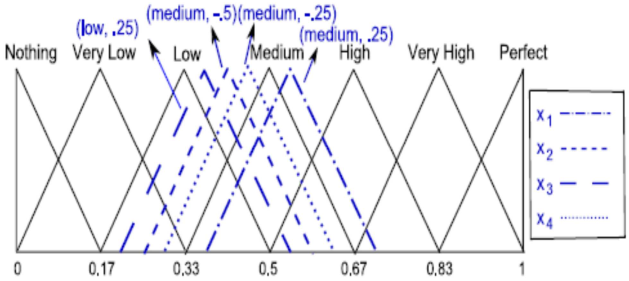

Fig. 4. Results from computational models based on 2-tuple linguistic representation.

The extended model solves the problem of loss of information that exists with the other two classic models. In this approach the computations performed are based on the extended model. This model can address a continuous valued attribute. It can easily make use of the computing with word approach thereby reducing the loss of information in other models. Finally, the result can always be expressed in the initial expression domain [36]. Since this model is followed in the fusion approach, the computations used in the representation of the model are mentioned with the following definitions.

4) Definitions: The linguistic information in the linguistic computational model is represented as a 2-tuple, *(*lt, *I),* where 'lt' is a linguistic term and *'I'* is a numerical value. The numerical value represents the symbolic translational value. Let a ∈ A be the set of numerical attributes listed in the dataset A and LT = {lt_0, ltg} be the linguistic term set with 'g' linguistic terms defined by the experts. The real values can be transformed into fuzzy set by means of a function T, as given in Eq. 3:

$$\mathrm{T(a)} = \{(\mathrm{lt}_0, \mathrm{u}_0), .., (\mathrm{lt}_\mathrm{g}, \mathrm{u}_\mathrm{g})\}, \ \mathrm{lt}_\mathrm{i} \in \mathrm{LT} \text{ and } \mathrm{U}_\mathrm{i} \in [0, 1], \\ \text{Such that } \mathrm{u}_\mathrm{i} = \mu_{\mathrm{lt}_\mathrm{i},}(a) \tag{3}$$

where T(a) = $\{(\mathrm{lt}_0, \mathrm{u}_0), .., (\mathrm{lt}_\mathrm{g}, \mathrm{u}_\mathrm{g})\}$ represents the transformation function given by set of 2-tuple pairs {($\mathrm{lt}_{0,}$ u_0), ($\mathrm{lt}_{\mathrm{g},}$ u_g)}. Here 'g' number of linguistic terms given by 'lt', is paired with symbolic translation value 'u'. $\mu_{\mathrm{lt}_\mathrm{i}}$ is the fuzzy membership function defined for linguistic term 'i' [3]. The membership function defined in the proposed approach is a triangular member function as defined in Eq 1. Definitions and functions used in this approach to compute the 2-tuple linguistic terms and the characteristic value associated with the real valued attribute are mentioned below [5]. The triangular fuzzy membership function μ_{lt} is defined as per the definition in Eq. 1 and the parameters are defined by the 3-tuple (a,b,c). From the fuzzy set a numerical value assessed in the interval [0, g] is obtained by a function χ as given by Eq. 4. β represents the numerical aggregated value.

$$= \beta \ \chi(\mathrm{T(a)}) = \chi\ ((\mathrm{ltj}, \mathrm{u}_\mathrm{j})\ \mathrm{j}= 0,...,\mathrm{g}) = \beta \\ \chi(\mathrm{T(a)}) = \frac{\sum_{j=0}^{g} ju(j)}{\sum_{j=0}^{g} u(j)} \tag{4}$$

Σ represents the summation operation. β is obtained by dividing the value obtained by the summation of the product of index j and the j $^{\mathrm{th}}$ symbolic value (u(j) or u_j) by the value obtained by the summation of the j $^{\mathrm{th}}$ symbolic value (u(j) or u_j). An approximate function is used to obtain the index of the result. From the obtained information, the linguistic 2-tuple values are generated using the following function Δ as shown in Eq. 5. This function is used to avoid any approximation process that may lead to loss of information. Here round (.) is the usual round operation. Linguistic term LT_i has the closest index label to β, and α is the value of the symbolic translation.

$$\Delta : [0, \mathrm{g}] \rightarrow \mathrm{LT} \text{ x } [-0.5, 0.5) \\ \Delta\ (\beta) = \ (\mathrm{LT}_\mathrm{t}, \alpha) \text{ where} \begin{cases} \mathrm{LT}_\mathrm{i}, & i = round\ (\beta) \\ \alpha = \ \beta - i, \alpha & \in [-0.5, 0.5] \end{cases} \tag{5}$$

After the numerical values and linguistic terms are aggregated, the information corresponding to each of the linguistic terms need to be generated [6]. The operation that is used for this purpose is the arithmetic mean value $\bar{x}$ given by Eq. 6.

$$\bar{x} = \ \sum \frac{1}{n} \sum_{i=1}^{n} \beta_i \tag{6}$$

During the re-computation phase of the real value, the functions ρ and η are used. The function ρ is used to compute, two 2-tuple information from the initial linguistic 2-tuple. For the linguistic term set with term LT, ρ is defined as given by Eq. 7.

$$\rho:[0, \mathrm{k}] \rightarrow \{\mathrm{LT} \times [0, 1]\} \times \{\mathrm{LT} \times [0, 1]\} \\ \rho\ (\beta) = \{(\mathrm{lt}_\mathrm{k}, 1 - \gamma), (\mathrm{lt}_{\mathrm{k+1}}, \gamma)\} \tag{7}$$

where k = trunc (β) and γ = β – k, trunc represents the usual truncation operation. β value is same as the value obtained in Eq.4. Let (lt_k, 1 − γ) and ($\mathrm{lt}_{\mathrm{k+1}}$, γ) be the computed two 2-tuples. The following function η is then used to compute the equivalent numerical value assessed in A as given by Eq. 8. A canonical characteristic value is computed using a function defined and it returns a characteristic value, CV (.).The function can be average, mean of max or as selected from the set of selected fuzzy operations [6].

$$\eta\ ((\mathrm{lt}_\mathrm{k}, 1 - \gamma)\ , (\mathrm{lt}_{\mathrm{k+1}}, \gamma)) = \mathrm{CV}\ (\mathrm{lt}_\mathrm{k})\ (1 - \gamma\ ) + \mathrm{CV}\ (\mathrm{lt}_{\mathrm{k+1}}) \\ \gamma \tag{8}$$

where CV (.) is a function providing a characteristic value.

5) Fuzzy Naive Bayes Classifier: A Naive Bayes Classifier is a widely used efficient classifier that applies Bayes' theorem with strong (naive) independence assumptions. It is a simple probabilistic classifier that handles both real and discrete data. The computation process is very easy and provides better speed and accuracy in classifier performance [38]. Storr proposed a fuzzy logic extension to the Naive Bayes Classifier with membership functions in applications with variables having continuous domain [26]. Membership function and fuzzy theory without loss of information were used in this approach. An aggregate fuzzy naive bayes classifier was proposed by Kayalp where the membership function was obtained using previous knowledge and 2-tuple linguistic knowledge approach and the function procedure followed the Chen's function principle [25].

C. Proposed Computational Framework

The three phases in the proposed approach are (i) The computation of real to fuzzified real value (ii) Re-computing the fuzzified real value in the initial expression domain (iii) Linguistic Fuzzy Naive Bayes Classifier (LFNBC) Model. A framework representing the process involved in the computations is shown in Fig. 5.

1) Computation of real to fuzzified real value: The first phase in the proposed approach was to select the numerical and categorical attributes. These attributes can be combined to generate linguistic terms with associated member function definitions. The support of the expert knowledge is sought to define the membership functions for the selected numerical features during this phase. This definition is based on the linguistic term set associated with the categorical feature. With expert knowledge, the features that contribute significantly to the outcome of the prediction are analyzed and defined. The defined member functions are used in the first phase of computing model to generate the fuzzified values. Major steps in computation of fuzzified real values are Unification, Transformation, and Fusion.

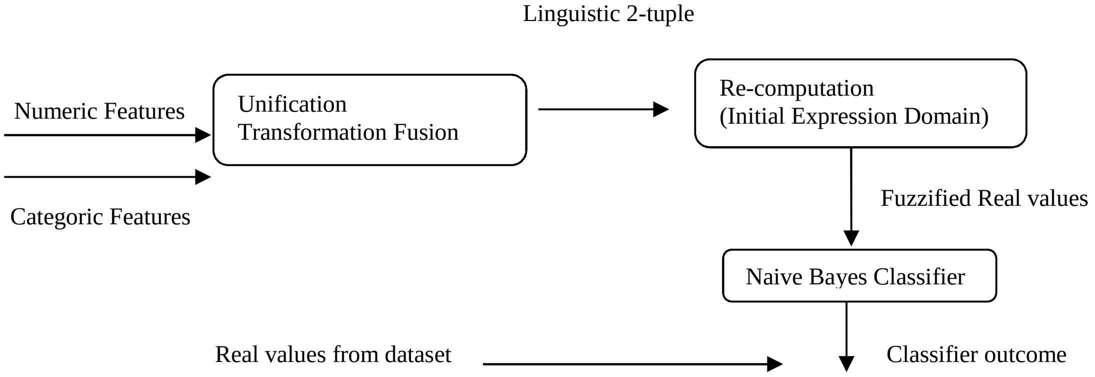

Fig. 5 Computational Framework of Linguistic Fuzzy Naive Bayes Classifier.

Unification into Fuzzy Sets (Normalization)–The information must be uniform in order to apply the fusion approach. The input information is unified using the concepts of fuzzy sets and linguistic approach. The usage of basic linguistic terms to normalize the information is processed during the first step. The linguistic terms should maintain the uncertainty degree and discriminate the expressions involving the performance values. The linguistic terms can be semantically same or different. Also, the number of terms varies largely with the attribute features and takes odd values ranging from 3 to 15 terms. Terms are defined in this step as shown in Eq. 2. Using the membership function definition in Eq. 1 the terms are defined and unified into fuzzy sets.

Transformation into 2-tuples–Due to complexity in computations with fuzzy sets, the information is transformed to 2-tuple linguistic fuzzy representation model. A function Xi that supports the information in the fuzzy set is defined. The computations result in generation of Beta value which can be easily transformed to 2–tuple linguistic value using another function Delta. Therefore, input information is unified and transformed to 2-tuple linguistic model after definitions of member function, Xi, Beta and Delta values. The computations to generate 2-tuples are performed in same order as given in Eq. 3, Eq. 4 and Eq. 5. The output obtained here is far away from the initial domain of expression. To maintain the comprehensibility of the data, a reconversion process is highly appreciated.

Fusion by 2-tuple fusion operator–A collective value of performance for each of the alternatives is generated by the fusion process. The value generated after the transformation step is in the form of 2-tuple linguistic representation. This information is aggregated to obtain the collective performance values of the alternatives, with a suitable fusion operator. The arithmetic means the operation is used to compute the collective information as given in Eq. 6.

2) Re-computing the fuzzified real value in the initial expression domain: The 2-tuple linguistic information is different from the information expressed in the initial domain of expression. In order to enhance the comprehensiveness of the information and the strategic decisions applied, the re-computational methods adopted are critical in the decision making an application. The fuzzified real values computed need to be converted back to the real domain of the expression to apply the standard classification techniques. Eq. 7 and Eq. 8 are used for the recomputation of real values which is also known as the “backward step” used in the computational step. The dataset after this phase of computation will have the selected numerical attributes, recomputed with the linguistic fuzzified values incorporated.

3) Linguistic fuzzy naive bayes classifier model: The Naive Bayes Classifier calculates the posterior probability by multiplying the probabilities determined along each attribute. In fuzzy classifiers we obtain a mapping from the attributes to the term sets. As mentioned in the above steps, these terms stand for fuzzy sets. The decision score again projects the belongingness of the data points to the membership functions defined. Integrating the fuzzy results with Naive Bayes Classifier yields better classifier results.

In the third phase of computation, the Linguistic Fuzzy Naïve Bayes Classifier is run on the real valued data set. The selected numerical attributes are defined using categorical terms of linguistic nature. The fusion approach is applied to these numerical-linguistic combination terms in the 2-tuple representation. Linguistic Fuzzy Naïve Bayes Classifier is run on the real values from the dataset, which also contain the recomputed fuzzified real values.

The numerical values obtained contain prior information. Fuzzy membership function segments the numeric data as per the function definition. In order to obtain a normalized distribution, the segmentation procedure is linked with the defined linguistic terms. The Naive Bayes Classifier is run using Eq. 9 for the real valued attributes.

$$LFNBC_{new}(C^*) = argmax \quad c_i \in c \left\{ p \left(\prod_{j=1}^{n} p(a_j \,|\, c_i) \right) \right\} \tag{9}$$

where C^* is the new class to be determined. Argmax is a function that returns the index of the maximum value, C_i represents all possible classes and a_j represents the attributes of the class.

III. Results and Discussion

The computations involving fuzzy sets, linguistic models and Naïve Bayes Classifier are done using ipython and weka tool. The heart disease dataset used for the experiments was obtained from UCI Machine Learning Repository [39]. The detailed description of the dataset, implementation details about the linguistic terms used the member function definitions, the computed aggregate performance values are given in the following subsections.

A. Dataset

The Statlog Heart Dataset is used for experimentation in the proposed approach. The information about the variables and the type and range of the values of the variables used in the experiment are provided in Table I. The dataset contains 270 instances with 13 features and one class attribute. For linguistic fuzzy approach method, 6 attributes [3 numeric and 3 linguistic] are selected from the listed features. The selected features from Table I are used for modelling.

B. Implementation Details

With expert knowledge advice the features contributing to the prediction of heart disease in patients and those applicable to the proposed model were selected for further experimental analysis. The parameters like blood pressure, cholesterol, chest pain, maximum heart rate show significant impact on the prediction of the heart disease. These parameters were selected for the linguistic fusion approach modelling. The study was conducted on 6 attributes, which include 3 categorical features and 3 numerical features that are defined using membership functions. The numerical features were aggregated to the linguistic terms for generating the 2-tuple linguistic model using fusion operators. The 3 numerical features selected were serum cholesterol (chol), resting blood pressure (rbp) and maximum heart rate (mhr). Each of them was aggregated by the linguistic terms resting electro cardiograph (recg), chest pain type (cpt) and slope (slp), respectively.

TABLE I
STATLOG HEART DATASET

Variable	Variable Definition	Type of Value	Range/Category of Values
AGE	Age of patient	Numeric	[29-77]
SEX	Gender of patient	Categorical	1-male; 0-female

CPT	Chest Pain Type	Categorical	1-typical angina;
			2-atypical angina;
			3-non-anginal pain;
			4-asymptomatic

RBP	Resting Blood Pressure	Numeric	[94-200]
SC	Serum Cholesterol	Numeric	[126,564]
FBS	Fasting Blood Sugar Value>120mg/dl	Categorical	[True, False]

RECG	Resting Electro Cardiograph	Categorical	0-normal;
			1-ST-Twave abnormality;
			2-LeftVentricular Hypertrophy

MHRA	Maximum Heart Rate Achieved	Numeric	[71-202]
EIA	Exercise Induced Angina	Categorical	[Yes, No]
STDep	ST Depression induced by exercise relative to rest	Numeric	[0-6.2]

Slope	Slope of peak exercise ST segment	Categorical	[Upslope;
			Flat;
			Downslope]

NFCMV	Number of fluoroscopies colored major vessels	Categorical	[0;1;2;3]

Thal-HSS	Heart Scan Status	Categorical	[Normal;
			Fixed defect;
			Reversible defect;
			Absent]

CVP	Class variable Prediction of Heart Disease	Categorical	Absence-0; Presence-1

The membership functions are defined for each of the Linguistic Terms (LT) in the term set. The term set for each of the linguistic term, LT_1, LT_2, LT_3 is defined as given in Eq.10. Term sets and their notations are explained in Table II. The definitions are based on the range or category of values mentioned in the dataset.

$$\begin{gathered} LT_1 \;=\; \{lt_0{=}NORM,\ lt1{=}ST\text{-}TWA,\ lt2{=}LVH\} \\ LT_2 = \{lt_0{=}TYPANG,\ lt1{=}ATYPANG,\ lt2{=}NONANG, \\ lt3{=}ASYMPT\} \end{gathered} \tag{10}$$

The upper and lower bounds of the member functions defined for the selected fuzzified attributes; the characteristic function value returned for each member function are also specified in Table II.

TABLE II
RANGE OF MEMBERSHIP FUNCTION VALUES AND CHARACTERISTIC VALUES

Term Sets	Membership Function [µlt(x)]	Symbolic Notation	Index Value	Lower Bound[a]	Characteristic Value[b]	Upper Bound[c]
Normal	NORM	*Nm*	0	126	290	354
ST-Twave abnormality	ST-TWA	*Stw*	1	197	269	327
Left Ventricular Hypertrophy	LVH	*Lvh*	2	149	390	564
Typical angina	TYPANG	*Tya*	0	94	175	180
Atypical angina	ATYPANG	*Atya*	1	94	174	192
Non-anginal pain	NONANG	*Nap*	2	100	173	200
Asymptomatic	ASYMPT	*Asy*	3	108	148	165
Upslope	USLOPE	*Usl*	0	96	149	202
Flat	FLAT	*Fl*	1	71	133	190
Downslope	DSLOPE	*Dsl*	2	96	145	194

The member functions are defined with the help of experts in the medical field. The characteristic values are computed using the average of the values representing the terms that fully belong to the member function. Setting the values for each category function is a crucial task that can only be accomplished with expert medical advice. Collective performance values are obtained using the paired features. The attribute preference values of all the selected attributes are given in Table III. The 2-tuple values obtained in symbolic translation is converted to values that represent information as membership degree. The values are generated for the first 20 instances in the dataset.

TABLE III
ATTRIBUTE PREFERENCE VALUES IN 2-TUPLES

Rbp	Cpt	Chol	Recg	Mhr	Slope
(Asy,-0.2)	(Asy,0)	(Lvh,1)	(Lvh,0)	(Fl,1)	(Fl,0)
(Tya,0.5)	(Nap,0)	(Lvh,2)	(Lvh,0)	(Fl,1)	(Fl,0)
(Tya,-0.3)	(Atya,0)	(Nm,1)	(Nm,0)	(Usl,1)	(Usl,0)
(Atya,-0.2)	(Asy,0)	(Nm,1)	(Nm,0)	(Fl,1)	(Fl,0)
(Atya,-0.3)	(Atya,0)	(Lvh,1)	(Lvh,0)	(Usl,1)	(Usl,0)
(Tya,-0.3)	(Asy,0)	(Nm,1)	(Nm,0)	(Usl,1)	(Usl,0)
(Atya,-0.2)	(Nap,0)	(Lvh,1)	(Lvh,0)	(Fl,1)	(Fl,0)
(Atya,-0.4)	(Asy,0)	(Lvh,1)	(Lvh,0)	(Fl,1)	(Fl,0)
(Nap,-0.3)	(Asy,0)	(Lvh,1)	(Lvh,0)	(Fl,1)	(Fl,0)
(Asy,-0.3)	(Asy,0)	(Lvh,2)	(Lvh,0)	(Fl,1)	(Fl,0)
(Tya,-0.3)	(Asy,0)	(Nm,1)	(Nm,0)	(Fl,1)	(Fl,0)
(Tya,-0.2)	(Asy,0)	(Lvh,1)	(Lvh,0)	(Usl,1)	(Usl,0)
(Tya,-0.2)	(Nap,0)	(Lvh,1)	(Lvh,0)	(Usl,1)	(Usl,0)
(Nap,0.4)	(Tya,0)	(Nm,1)	(Nm,0)	(Fl,1)	(Fl,0)
(Atya,-0.3)	(Asy,0)	(Lvh,1)	(Lvh,0)	(Usl,1)	(Usl,0)
(Tya,-0.3)	(Asy,0)	(Nm,0)	(Nm,0)	(Fl,1)	(Fl,0)
(Nap,0.2)	(Asy,0)	(Nm,1)	(Nm,0)	(Fl,1)	(Fl,0)
(Tya,-0.3)	(Asy,0)	(Lvh,1)	(Lvh,0)	(Dsl,1)	(Dsl,0)
(Tya,-0.3)	(Tya,0)	(Lvh,1)	(Lvh,0)	(Fl,1)	(Fl,0)
(Tya,-0.2	(Tya,0)	(Nm,1	(Nm,0)	(Usl,1)	(Usl,0)

The arithmetic mean values are computed from the 2-tuple values by symbolic translation. It represents the information as membership degree. The membership degree obtained by this method represents a more accurate value that combines the numeric and linguistic feature values. The arithmetic mean value for each of the three paired linguistic term set for the 20 instances in the dataset is as given in Table IV. The LFNBC uses this value and produces a more accurate performance compared to the SNBC.

C. Result Analysis

From the Statlog Heart dataset 243 instances are randomly chosen to form the training set and 27 instances are used as test set. The results obtained showed 77.7% accuracy for Simple Naïve Bayes Classifier (SNBC) and 91.3% accuracy for the proposed Linguistic Fuzzy Naïve Bayes Classifier (LFNBC).

TABLE IV
ARITHMETIC MEAN VALUES

Rbp	**Cpt**	**Mean β (rbp-cpt)**	**Chol**	**Recg**	**Mean β (chol-ecg)**	**Mhr**	**Slp**	**Mean β (mhr-slp)**
130	Asympt	2.4	322	Lvh	.5	109	Flat	1
115	Nonang	1.8	564	Lvh	2	160	Flat	1
124	Atypang	1.4	261	Norm	0.5	141	Uslope	0.5
128	Asympt	2.4	263	Norm	1	105	Flat	1
120	Atypang	1.3	269	Lvh	0.5	121	Uslope	0.5
120	Asympt	2.3	177	Norm	0.5	140	Uslope	0.5
130	Nonang	1.9	256	Lvh	1	142	Flat	1
110	Asympt	2.1	239	Lvh	1	142	Flat	1
140	Asympt	2.4	293	Lvh	1	170	Flat	1
150	Asympt	2.3	407	Lvh	1	154	Flat	1
135	Asympt	2.4	234	Norm	1	161	Flat	1
142	Asympt	2.3	226	Lvh	0.5	111	Uslope	0.5
140	Nonang	1.9	235	Lvh	0.5	180	Uslope	0.5
134	Typang	0.9	234	Norm	1	145	Flat	1
128	Asympt	2.4	303	Lvh	0.5	159	Uslope	0.5
112	Asympt	2.2	149	Norm	1	125	Flat	1
140	Asympt	2.4	311	Norm	1	120	Flat	1
140	Asympt	2.4	203	Lvh	1.5	155	Dslope	1.5
110	Typang	0.6	211	Lvh	1	144	Flat	1
140	Typang	0.9	199	Norm	0.5	178	Uslope	0.5

For each method the classification accuracy (ratio of correctly classified cases of presence and absence of heart disease), true positive rate (the proportion of actual diseased cases which are correctly identified as such), true negative rate (proportion of cases of no heart disease that are correctly identified as absence of disease) are analysed. Classification accuracy refers to the ratio of the number of correctly classified cases and is equal to the sum of True Positive (TP) and True Negative (TN) divided by the total number of cases N. Sensitivity refers to the rate of correctly classified positive (True Positive Rate) and is equal to the ratio of patients with presence of heart disease who are accurately considered as the ones with the disease and is computed as TP to the sum of TP and False Negative (FN). Specificity refers to the ratio of patients who have no heart disease and who are accurately considered as patients without heart disease. It is the rate of correctly classified negative (True Negative Rate) and is equal to TN divided by sum of TN and False Positive (FP). The performance obtained by the Simple Naïve Bayes Classifier and the proposed Linguistic Fuzzy Naïve Bayes Classifier are summarized in Table V.

TABLE V
PERFORMANCE COMPARISON OF CLASSIFIERS (SNBC-LFNBC)

Classifier	Accuracy (%)	Sensitivity (%)	Specificity (%)
LFNBC	91.3	92.68	90.19
SNBC	83.6	88.63	79.1

The three statistical measures, classification accuracy, sensitivity and specificity are used to evaluate the performance of each of the classification model under study. The graphical representation of the performance comparison is shown in Fig.6.

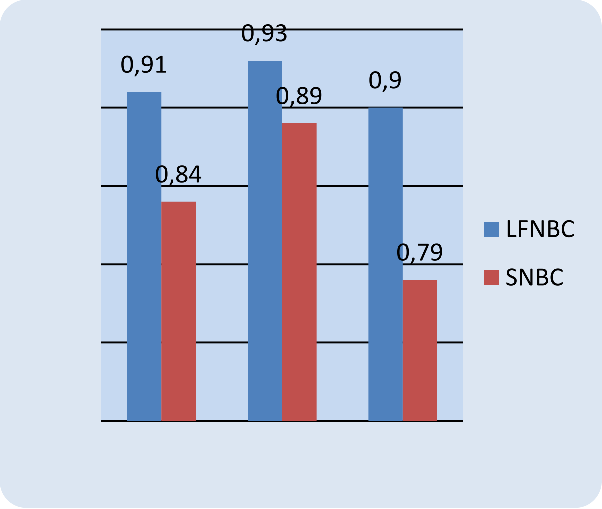

Fig. 6. Performance comparison of classifiers SNBC-LFNBC

IV. CONCLUSIONS

The proposed Linguistic Fuzzy Naive Bayes Classifier is an extended version of Simple Naive Bayes Classifier and Fuzzy Naive Bayes Classifier. The approach of fuzzifying and re-computing the real numerical values in the initial expression domain, gives the classifier the flexibility of combining the real values in the dataset for classification purpose. The approach has the convenience of not using the fuzzy computations with extension principle. Using extension principle would give rise to computational complexities with loss of information. Another approach in literature uses Chens function principle for computations. This principle was based on selection of equal number of numeric and linguistic terms for combination into 2-tuple representation. The computations were performed on even number of attributes. The proposed approach overcomes these limitations by recomputing the real values from the fuzzified values. The approach enables the integration of fuzzy techniques with the classification algorithm. Apart from the recomputed fuzzified real values, the real values from the dataset are also included in the classifier execution. This inclusion removes the limitation of running only even number of attributes in the classifier. As an extension to the model proposed, the attribute preferences can be modified by adding weights to the selected attributes. The method can be applied to more attributes in the dataset after seeking advice from the experts in the knowledge domain. Various decision-making applications in the field of banking, information retrieval and supplier selection can make use of this approach to bring in more accurate predictions.

REFERENCES

[1] J. Han, M. Kamber, and J. Pei, “Data Mining: Concepts and Techniques,” Third. Waltham, USA: Elsevier Inc., 2012.

[2] R. J. Roiger, “Data Mining A Tutorial-Based Primer,” 2nd ed. London: CRC Press, Taylor & Francis Group, 2017.

[3] L.A. Zadeh, “The concept of a linguistic variable and its applications to approximate reasoning,” *Information Sciences*, vol.8, pp. 199–249, 1975.

[4] L.A. Zadeh, “Fuzzy Logic = Computing with Words,” *IEEE Transactions on Fuzzy Systems*, vol.4, pp. 103–111, 1996.

[5] F. Herrera and L. Martinez, “An approach for combining linguistic and numerical information based on 2-tuple fuzzy linguistic representation model in decision-making,” *International Journal of Uncertainty, Fuzziness, Knowledge-Based Systems*, vol. 8, pp. 539–562, 2000.

[6] F. Herrera, E. Herrera-Viedma, and L. Martínez, “A fusion approach for managing multi-granularity linguistic terms sets in decision making,” *Fuzzy Sets Systems*., vol. 114, pp. 43–58, 2000.

[7] C. C. Li, Y. C. Dong, F. Herrera, E. Herrera-Viedma, and L. Martinez, “Personalised individual semantics in computing with words for supporting linguistic group decision making. An application on consensus reaching,” *information Fusion*, vol. 33, pp. 29-40, 2017.

[8] R. Urena, F. Chiclana, J. A. Morente-Molinera, and E. Herrera-Viedma, “Managing incomplete preference relations in decision making: a review and future trends,” *Information Sciences*, vol. 302, pp. 14–32, 2015.

[9] R. Urena, F. Chiclana, and E. Herrera-Viedma, “Consistency based completion approaches of incomplete preference relations in uncertain decision contexts,” *in: IEEE International Conference on Fuzzy Systems, 2015*, pp. 1–6.

[10] M. Brunelli, and M. Fedrizzi, “Boundary properties of the inconsistency of pair wise comparisons in group decisions,” *European Journal of Operational Research*, vol. 240, pp. 765–773, 2015.

[11] S. Kubler, W. Derigent, A. Voisin, J. Robert, and Y. Le Traon, “Knowledge-based Consistency Index for Fuzzy Pairwise Comparison Matrices,” *in: IEEE International Conference on Fuzzy Systems*, 2017, pp.1–7.

[12] R. Verma and B. D. Sharma, “Trapezoid Fuzzy Linguistic Prioritized Weighted Average Operators and Their Application to Multiple Attribute Group Decision Making,” *Journal of Uncertainty Analysis and Applications*, vol. 2, pp. 1-19, 2014.

[13] R. Verma, “Multiple Attribute Group Decision Making based on Generalized Trapezoid Fuzzy Linguistic Prioritized Weighted Average Operator,” *International Journal of Machine Learning and Cybernetics*, 2016

[14] R. Verma, “Prioritised Information Fusion Method for Triangular Fuzzy Information and Its Application to Multiple Attribute Decision Making,” *International Journal of Uncertainty Fuzziness and Knowledge Based Systems*, vol. 24, no. 2, pp. 265-289, 2016.

[15] I. Beg. and T. Rashid, “An intuitionistic 2-tuple linguistic information model and aggregation operators”, *International Journal of Intelligent Systems*, Vol. 31 No. 6, pp. 569-592, 2016.

[16] H. Liao, X. Mi, Z. Xu, J. Xu, and F. Herrera, "Intuitionistic Fuzzy Analytic Network Process," in *IEEE Transactions on Fuzzy Systems*, vol.26, no.5, pp. 2578-2590, 2018.

[17] S. M. Chen. and W. H. Han, “A new multiattribute decision making method based on multiplication operations of interval-valued intuitionistic fuzzy values and linear programming methodology,” *Information Sciences*, vol. 429, no. 2, pp. 421-432, 2018.

[18] S.H. Cheng, “Autocratic multiattribute group decision making for hotel location selection based on interval-valued intuitionistic fuzzy sets,” Information *Sciences*, vol. 427, no. 1, pp. 77-87, 2017.

[19] J. Deepa, and S. Kumar, “Improved accuracy function for interval-valued intuitionistic fuzzy sets and its application to multi-attributes group decision making,” *Cybernetics and Systems*, vol. 49, no. 1, pp. 64-76, 2018.

[20] P. D. Liu, and S. M. Chen, “Multiattribute group decision making based on intuitionistic 2-tuple linguistic information,” *Information Sciences*, vol. 430-431, no. 1, pp. 599-619, 2018.

[21] P. Wang, X. H. Xu, J. Q. Wang, and C. G. Cai, "Interval-valued intuitionistic linguistic multi-criteria group decision-making method based on the interval 2-tuple linguistic information", *Journal of Intelligent & Fuzzy Systems,* vol. 33, no. 2, pp. 985-994, 2017.

[22] F. T. Bahari, and M. S. Elayidom, "An Enhanced Analytic CRM Framework Using Symbolic Fuzzy Approach in Decision Making Applications", *Journal of Advanced Research in Dynamical & Control Systems,* vol. 10, 15-Special Issue, 2018.

[23] F. Chiclana, F. Mata, L. G. Perez, and E. H. Viedma, "Type-1 OWA Unbalanced Fuzzy Linguistic Aggregation Methodology: Application to Eurobonds Credit Risk Evaluation," *International Journal of Intelligent Systems,* vol.33, pp.1071-1088, 2018.

[24] A. Peña, I. Bonet, C. Lochmuller, F. Chiclana, and M. Góngora, "An integrated inverse adaptive neural fuzzy system with Monte-Carlo sampling method for operational risk management," *Expert Systems with Applications*, vol. 98, pp.11-26, 2018.

[25] G. Yazgı Tütüncu, and Necla Kayaalp, "An aggregated fuzzy naive bayes data classifier," *Journal of Computational and Applied Mathematics*, vol. 286, pp. 17–27, 2015.

[26] H. P. Störr, "A compact fuzzy extension of the naive bayesian classification algorithm," *Proceedings of the Third International Conference on Intelligent Technologies and Vietnam-Japan Symposium on Fuzzy Systems and Applications*, pp.172-177, 2002.

[27] Y. Z. Xi, W. T. Xue, and S. L. Joon, "Fuzzy naive bayesian for constructing regulated network with weights," *Bio-Medical Materials and Engineering, v*ol. 26, pp. S1757–S1762, 2015.

[28] Y. Tang, W. Pan, H. Li, and Y. Xu, "Fuzzy naive bayes classifier based on fuzzy clustering", *Proceedings of 2002 IEEE International Conference on System, Man and Cybernetics*, 2002.

[29] Y. Tang, and Y. Xu, "Application of fuzzy naive bayes and a real-valued genetic algorithm in identification of fuzzy model, *Information Sciences*, v.169, p.205-225, 2005.

[30] S. Palaniappan, and R. Awang, "Intelligent heart disease prediction system using data mining techniques," *IEEE/ACS International Conference on Computer Systems and Applications*, 2008.

[31] S. B. Patil, and Y. S. Kumaraswamy, "Extraction of significant patterns from heart disease warehouses for heart attack prediction," *International Journal of Computer Science and Network Security*, vol. 9, pp. 228–235, 2009.

[32] Jan Bohacik, and Michal Zabovsky, "Naive bayes for statlog heart database with consideration of data specifics," *IEEE 14th International Scientific Conference on Informatics*, 2017.

[33] O. W. Samuel, G. M. Asogbon, A. K. Sangaiah, P. Fang, and G. Li, "An integrated decision support system based on ANN and Fuzzy AHP for heart failure risk prediction," *Expert Systems with Applications*, vol. 68, pp.163−172, 2017.

[34] K. Uyar, A. Ilhan, "Diagnosis of heart disease using genetic algorithm based trained recurrent fuzzy neural networks," *Procedia Computer Science*, vol. 120, pp.588-593, 2017.

[35] G. T. Reddy, N. Khare, "An Efficient System for Heart Disease Prediction using Hybrid OFBAT with Rule-Based Fuzzy Logic Model," *Journal of Circuits, Systems, and Computers*, vol.26, 2017

[36] L. Martinez, D. Ruan, F. Herrera, E. Herrera-Viedma, and P.P. Wang, "Linguistic decision making: tools and applications," *Information Sciences*, vol.179, pp. 2297–2298, 2009.

[37] L. Martinez, "Computing with words in linguistic decision making: Analysis of linguistic computing models," *IEEE International Conference on Intelligent Systems and Knowledge Engineering,* 2010.

[38] I. H. Witten, E. Frank, and M. A. Hall, "Practical *Machine Learning Tools and Techniques,*" 3rd Edition. Burlington, MA, USA: Morgan Kaufman Publishers, 2011.

[39] UCI Repository of Machine Learning Databases, http://archive.ics.uci.edu/ml/datasets/Statlog+%28Heart%29.