Assessment of Fuzzy Gaussian Naive Bayes for Classification Tasks

Jodavid A. Ferreira, Elaine A. M. G. Soares, Liliane S. Machado and Ronei M. Moraes

Laboratory of Statistics Applied to Image Processing and Geoprocessing
Federal University of Paraiba
João Pessoa, Brazil
Email: jodavid.arts@gmail.com, elaine.soares@ci.ufpb.br,
liliane@di.ufpb.br, ronei@de.ufpb.br

***Abstract*—Statistical methods have been used in order to classify data from random samples. In general, if we know the statistical distribution of data, we can use specific classifiers designed for that distribution and expect good results. This work assesses the accuracy of a Fuzzy Gaussian Naive Bayes (FGNB) classifier for tasks using data from five different statistical distributions: Negative Binomial, Logistic, Log-Normal, Weibull and Gamma. The FGNB classifier was recently proposed as a fuzzy extension of Gaussian Naive Bayes for training assessment in virtual environments. Results of assessment are provided and show different accuracy according to the statistical distribution of data.**

Keywords–Fuzzy Gaussian Naive Bayes Classifier, Classification, Accuracy Assessment.

I. INTRODUCTION

Statistical methods have been widely used in order to classify data from random samples [1]. In general, if we know the statistical distribution of data, we can use specific classifiers designed for that distribution and we can expected good results from that use [2]. Classifiers based on Gaussian distribution were exhaustively studied in the literature [3][4] and applied in several kinds of problems [5][6][7]. Some of their variations are known as Classical Bayes Rule [8] and Gaussian Naive Bayes [9].

However, in several kind of applications, it is not possible to affirm the sample data were measured with accuracy. In these cases, the imprecision on data should be incorporated in the classification method. Nowadays, a possible approach for this modelling is using fuzzy sets proposed by Lofti A. Zadeh [10]. Several classification methods based on fuzzy sets can be found in the literature and some of them are based on probability measures of fuzzy events [11]. Among them, the Fuzzy Gaussian Naive Bayes (FGNB) method was proposed by Moraes and Machado [12] and has been applied to classification and training assessment problems [13][14][15] achieving good results.

The main question about the classifiers based on Gaussian distribution is related to some classification problems in which data did not follow Gaussian distribution. So, it is interesting to know the limitations when those methods are used. This paper aims to verify if the FGNB method has good performance when classifying data given by the Logistic, Gamma, Weibull, Log-Normal and Negative Binomial distributions. For each statistical distribution were used data dimensions from 1 to 4.

The FGNB classifier was proposed recently; then it is necessary to know its accuracy. A preliminary performance analysis from these authors using FGNB classifier and other statistical distributions was performed [15]. In this paper, we enlarge the ¿ range of distributions used to verify the accuracy of the method. The results of those comparisons are analysed with respect to the better statistical distribution of data to be used for better FGNB performance, according to each dimension of data.

Section 2 presents the Fuzzy Gaussian Naive Bayes (FGNB) classification method which was used for data classification. In Section 3, the methodological part is described: the data used and how the samples were generated. In Section 4, the classification results for the 5 distributions statistics are detailed. The conclusion of the study, highlighting the distribution that was not well sorted, is in Section 5.

II. FUZZY GAUSSIAN NAIVE BAYES (FGNB)

Formally, let the classes of performance in space of decision be $\Omega = \{1, ..., M\}$ where M is the total number of classes. Let X be a vector of training data, according to sample data D, where X is a vector with n distinct features, **i.e.,** $X = \{X_1, X_2, , X_n\}$ and w_i, $i \in \Omega$ is the class in space of decision for the vector X. So, the probability of the class w_i, given the vector X, can be estimated using the Bayes Theorem:

$$P(w_i|X) = \frac{P(X|w_i)P(w_i)}{P(X)} = \frac{[P(X_1, X_2, \ldots, X_n|w_i)P(w_i)]}{P(X)} \quad (1)$$

Let us assume a naive hypothesis, in which each feature X_k is conditionally independent of every other feature X_l , for all $k \neq l \leq n$. This hypothesis, though sometimes it is not exactly realistic, enables an easier calculation of (1). An advantage of that assumption is the robustness acquired by classifier that now can classify data for which it was not trained for [16]. So, unless a scale factor S, which depends on $X_1, X_2, \ldots, X_n$, the equation (1) can be expressed by:

$$P(w_i|X_1, X_2, \ldots, X_n) = \frac{P(w_i)}{S} \prod_{k=1}^{n} P(X_k|w_i) \quad (2)$$

A possible approach is to assume Gaussian distribution for X and compute its parameters from D, i.e., mean vector

and covariance matrix [17]. From equation (2) it is possible to use the logarithm function in order to simplify the exponential function in the Gaussian distribution formula and, consequently, to reduce computational complexity replacing multiplications by additions:

$$g(w_i, X_1, X_2, \ldots, X_n) = log[P(w_i|X_1, X_2, \ldots, X_n)] = \quad (3)$$
$$= log\frac{P(w_i)}{S} + \sum_{k=1}^{n} log[P(X_k|w_i)]$$

where g is the classification function.

At this point, it is assumed that random variables $X_1, X_2, \ldots, X_n$ are also fuzzy variables because we are going to use their membership functions $\mu_{wi}(X_k)$ for this calculus [18]. Then, based on probability of a fuzzy event [11], the **FGNB** is done by [12]:

$$g_f(w_i, X_1, X_2, \ldots, X_n) = log[P(w_i|X_1, X_2, \ldots, X_n)] = \quad (4)$$
$$log\frac{P(w_i)}{S_f} + \sum_{k=1}^{n} log\mu_{wi}(X_k)P(X_k|w_i)$$

where g_f is the new classification function and $S_f$is a new scale factor.

The necessary parameters to compute $P(X_k|w_i)$ and $\mu_{wi}(X_k)$ should be learned from sample data D. The better estimation for class of the vector X can be obtained from the highest values of the classification function g_f. However, as S_f is a scale factor, it is not necessary to compute it for this maximization process. Then:

$$X \in w_i logP(w_i) + \sum_{k=1}^{n} log[\mu_{wi}(X_k)P(X_k|w_i)] > \quad (5)$$
$$logP(w_j) + \sum_{k=1}^{n} log[\mu_{wj}(X_k)P(X_k|w_j)]$$

is the classification rule for FGNB.

III. ASSESSMENT METHODOLOGY

Several studies show that assessment methods present better results when they are applied with data from a particular statistical distribution. In general, each method can achieve better results when data follow some specific statistical distributions [14]. In a previous work, Moraes [14] studied the FGNB method for classification tasks using six different statistical distributions: Binomial, Continuous and Discrete Uniform, Exponential, Gaussian and Poisson [15]. However, since the FGNB is a recent method, its performance is not clear with this other five statistical distributions: Negative Binomial, Logistic, Log-Normal, Gamma and Weibull. In this paper, we use the Monte Carlo simulation [19] to investigate the behaviour of this method.

For our implementation, the samples were generated with Monte Carlo simulation for 1, 2, 3, 4 and dimensions for the five distributions in two different formats. One is used for training the FGNB method and the other for testing the method. Their settings obey the following rules:

a) Random training sample: used for the training of the method, this sample has 40000 observations for all the 4 classes.

b) Random test sample: after the training, the method used this to the assessment. This sample was composed by 30000 observations for each class, totaling 120000 observations.

The assessment method FGNB was implemented to all the variety of dimensions and the respective classification matrices can be stored in order to assess the accuracy of this methodology in practical applications. In particular, we had used FGNB as a kernel of an online assessment method of virtual reality simulators for training [12][13].

A. SIMULATION

To use the method, random samples were generated for the 5 statistical distributions. The samples were generated in *software R* [20] using the following parameters for each distribution:

1) Negative Binomial: For the Binomial distribution denoted by $X \sim BN(p, k)$, 2 parameters are necessary for samples generation. The parameters used were:

TABLE I. PARAMETERS OF THE NEGATIVE BINOMIAL DISTRIBUTION.

NEGATIVE BINOMIAL $X \sim BN(p,k)$	CLASS 1	CLASS 2	CLASS 3	CLASS 4
DIMENSION 1	(0.4,10)	(0.4,30)	(0.2,30)	(0.4,130)
DIMENSION 2	(0.6,10)	(0.3,30)	(0.4,20)	(0.5,140)
DIMENSION 3	(0.4,30)	(0.4,10)	(0.4,130)	(0.3,47)
DIMENSION 4	(0.3,10)	(0.4,80)	(70,0.5)	(0.4,130)

2) Logistic: Samples were generated from the logistics distribution, using the following parameters:

TABLE II. PARAMETERS OF THE LOGISTIC DISTRIBUTION.

LOGISTIC $X \sim L(\mu,\sigma)$	CLASS 1	CLASS 2	CLASS 3	CLASS 4
DIMENSION 1	(0,2)	(20,2.5)	(43,2)	(60,3)
DIMENSION 2	(13,3)	(60,4)	(35,2)	(90,4)
DIMENSION 3	(20,3)	(40,2)	(108,4)	(72,4)
DIMENSION 4	(79,5)	(6,3)	(110,2)	(40,4)

3) Log-Normal: For the generation of Log-Normal distribution samples, the following parameters were used:

TABLE III. PARAMETERS OF THE LOG-NORMAL DISTRIBUTION.

LOG-NORMAL (μ,σ)	CLASS 1	CLASS 2	CLASS 3	CLASS 4
DIMENSION 1	(2,0.2)	(3,0.3)	(3.8,0.3)	(4.5,0.2)
DIMENSION 2	(2.8,0.2)	(2,0.3)	(4.7,0.2)	(3.7,0.3)
DIMENSION 3	(2,0.3)	(2.65,0.2)	(3.7,0.3)	(4.5,0.2)
DIMENSION 4	(4.2,0.2)	(3.5,0.1)	(2,0.4)	(3,0.3)

4) Gamma: The gamma distribution has the following notation $X \sim Gamma(shape, scale)$ and the parameters used for generation of the samples were:

 ISBN: 978-1-61208-393-3

TABLE IV. PARAMETERS OF THE GAMMA DISTRIBUTION.

GAMMA $(shape, scale)$	CLASS 1	CLASS 2	CLASS 3	CLASS 4
DIMENSION 1	(20,0.25)	(40,0.25)	(60,0.25)	(90,0.25)
DIMENSION 2	(12,1.0)	(32,1.0)	(65,1.0)	(110,1.0)
DIMENSION 3	(50,0.33)	(80,0.33)	(120,0.33)	(170,0.33)
DIMENSION 4	(80,0.17)	(130,0.17)	(190,0.17)	(250,0.17)

5) Weibull: The Weibull distribution has two parameters called shaped parameter and scale parameter. The parameters used to generate the samples were:

TABLE V. PARAMETERS OF THE WEIBULL DISTRIBUTION.

WEIBULL $(shape, scale)$	CLASS 1	CLASS 2	CLASS 3	CLASS 4
DIMENSION 1	(50,5)	(100,10)	(150,20)	(200,20)
DIMENSION 2	(50,5)	(100,10)	(150,20)	(200,20)
DIMENSION 3	(50,5)	(15,20)	(100,10)	(200,20)
DIMENSION 4	(200,20)	(50,5)	(150,20)	(100,10)

B. KAPPA COEFFICIENT

The Kappa Coefficient K proposed by Cohen [21] is a robust pondered measure which takes into account agreements and disagreements between two sources of information from a classification matrix [22]:

$$K = \frac{(P_0 - P_c)}{(1 - P_c)} \tag{6}$$

where: $P_0 = \frac{\sum_{i=1}^{M} n_{ii}}{N}$ and $P_c = \frac{\sum_{i=1}^{M} n_{i+} n_{+i}}{N^2}$, where n_{ii} is the total of main diagonal in the classification matrix; n_{i+} is the total of line i in this matrix, n_{+i} is the total of colum in the same matrix, M is the total number of classes and N is the total number of possible decisions presented in the matrix.

The variance of Kappa Coefficient K, denoted by σ_K^2 , is done by:

$$\sigma_K^2 = \theta_1 + \theta_2 + \theta_3. \tag{7}$$

where θ_1 , θ_2 , and θ_3 are given by:

$$\theta_1 = \frac{P_0(1 - P_0)}{N(1 - P_c)^2} \tag{8}$$

$$\theta_2 = \frac{2(1 - P_0) + 2P_0P_c - \theta_4}{N(1 - P_c)^3} \tag{9}$$

$$\theta_3 = \frac{(1 - P_0)^2\theta_5 - 4P_c^2}{N(1 - P_c)^4} \tag{10}$$

and the parameters θ_4 and θ_5 are done by:

$$\theta_4 = \frac{\sum_{i=1}^{M} n_{ii}(n_{i+} + n_{+i})}{N^2} \tag{11}$$

$$\theta_5 = \frac{\sum_{i=1}^{M} n_{ii}(n_{i+} + n_{+i})^2}{N^3} \tag{12}$$

An approximation to the first component of 7 can be used for calculations. However, in this paper, we used the complete formula for variance and the Kappa Coefficient was computed for the best results. This coefficient is widely used in the literature of pattern classification [2].

According to Landis and Koch, the Kappa coefficient can be interpreted as presented in Table VI [23]. By these interpretation it is possible to distinguish where lies the classification data.

TABLE VI. CLASSIFICATION OF KAPPA COEFFICIENT.

Kappa Statistic	Strength of Agreement
$<$0.00	Poor
0.00-0.20	Slight
0.21-0.40	Fair
0.42-0.60	Moderate
0.61-0.80	Substantial
0.81-1.00	Almost Perfect

IV. RESULTS

A. Negative Binomial Distribution

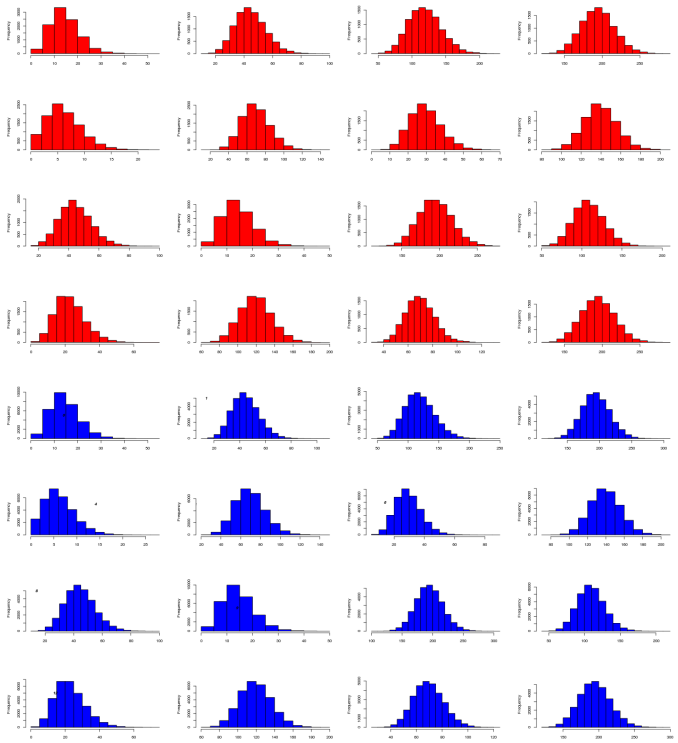

Figure 1. Random numbers generated for Negative Binomial distribution four sets of training and test samples (lines) for four classes (columns).

The negative binomial distribution is a discrete distribution, in which are considered some conditions: the experiment consists on an undetermined amount of repeated attempts, the probability of success is the same in each trial and the trials are independent. Using the method FGNB on Negative Binomial distribution with one dimension, was obtained a percentage accuracy of 69.91% and Kappa coefficient of 59.88% with a variance of 9.29×10^{-6} . To dimension of size 2, the percentage of correct answers was 93.94%, the Kappa cefficient 91.92% and variance 2.53×10^{-6}. With the dimension equal to 3, the percentage accuracy and Kappa coefficient were greater than

99% and the variance obtained was 2.62×10^{-7}. The best results were obtained with the data dimension equal to 4: the percentage of right classifications was 99.9%, the Kappa 99.8% and the coefficient variance 6.10×10^{-8}. The histograms of the samples for this distribution are available in Figure 1, where the red and blue ones represent the training sample and the test samples, respectively. In each histogram, we represent the dimensions on the lines and the classes on the columns.

B. Logistic Distribution

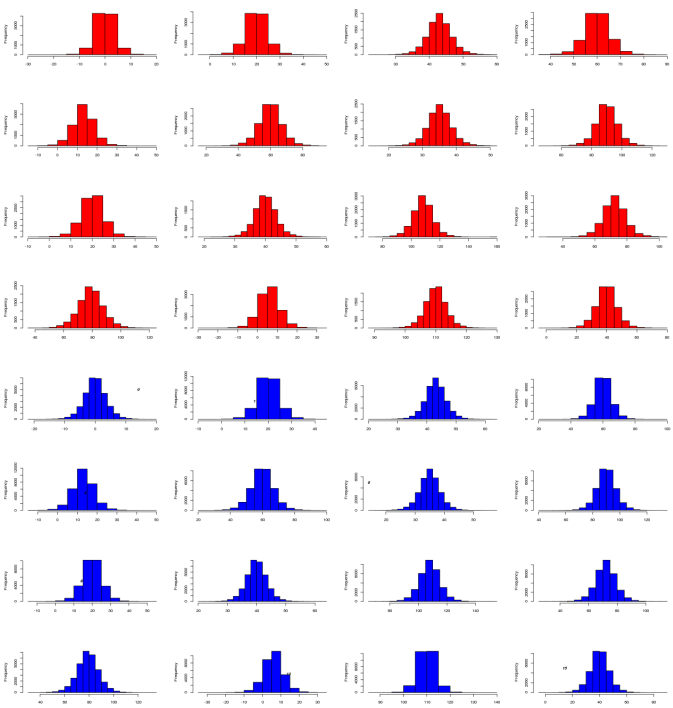

Figure 2. Random numbers generated for Logistic distribution four sets of training and test samples (lines) for four classes (columns).

The logistics distribution is a continuous distribution used in studies of population growth and agricultural production. It is also used in replacement of normal distribution due to the shape similarity of them in some specific studies. Figure 2 shows samples from logistics distribution used in this study. For a dimension equal to 1, the method proved to be more efficient for the logistic distribution than to the negative binomial distribution. The percentage accuracy was 84.41% and the Kappa coefficient was 79.22% with a variance of 1.94×10^{-6}. With dimension 2, the percentage of right classifications was 98.49%, with Kappa coefficient 97.99% and variance 2.19×10^{-7}.

The percentage of right classifications reached 99.9% when the point size was 3 under these conditions the Kappa coefficient was 99.5% and the variance 1.64×10^{-8}. For a dimension equal to 4, the method for distribution logistics FGNB achieved 99.9% of right classifications with a Kappa coefficient of 99.7% and variance 3.45×10^{-9}.

C. Log-Normal Distribution

The Log-Normal distribution is continuous and can be used to feature the lifetime of products and materials (semi-

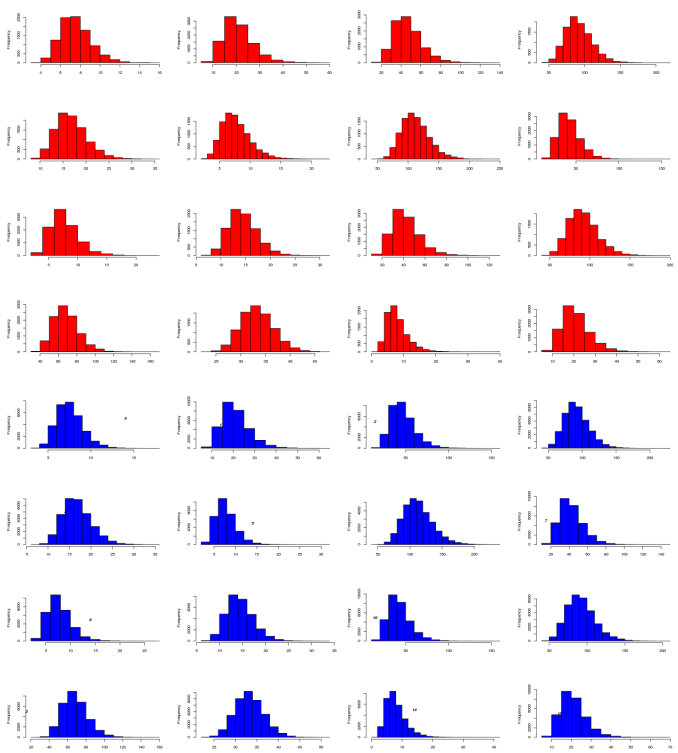

Figure 3. Random numbers generated for Log-Normal distribution four sets of training and test samples (lines) for four classes (columns).

conductors, diodes and electrical insulation, among others). The histograms of random numbers generated for simulations using Log-Normal distribution are presented in Figure 3. In a simulation with only one dimensional data, the Kappa coefficient was 67.69% with variance $2.71 \times 10^{-}6$ and 29074 misclassifications. With two dimensions was obtained Kappa equal to 83.35% with $1.61 \times 10^{-}6$ of variance. When using 3 dimensions, the results were 84.62% and $1.50 \times 10^{-}6$ for Kappa and its variation. And with 4 dimensions, 13513 misclassifications occurred and the kappa was 84.98% with $1.47 \times 10^{-}6$.

D. Gamma Distribution

The Gamma distribution is a continuous probability distribution, which has two parameters, the first one for shape and the second one for scale, and it requires that both parameters are greater than zero. The use of the FGNB method on the samples of the Gamma distribution that are present in Figure 4 produced the following results: 1 dimension, the percentage of right classifications was 72.33%, with a Kappa coefficient of 63.11% and variance of 2.92×10^{-06}. For dimension 2, the percentage of right classifications was 87.96%, with Kappa coefficient 83.96% and variance 1.57×10^{-06}. The percentage accuracy, for 3 dimensions, was greater than 97%, the Kappa coefficient 97.06% and the variance 3.199×10^{-07}. With dimension equal to 4, the kappa coefficient was 99.8% with a variance of 1.89×10^{-08} and the percentage of right classifications 99.87% with 153 misclassifications.

 ISBN: 978-1-61208-393-3

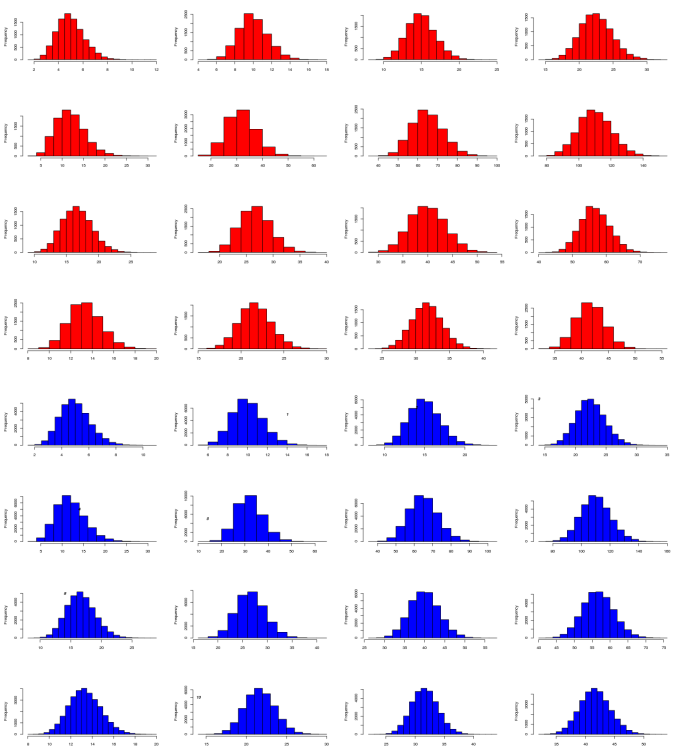

Figure 4. Random numbers generated for Gamma distribution four sets of training and test samples (lines) for four classes (columns).

E. Weibull Distribution

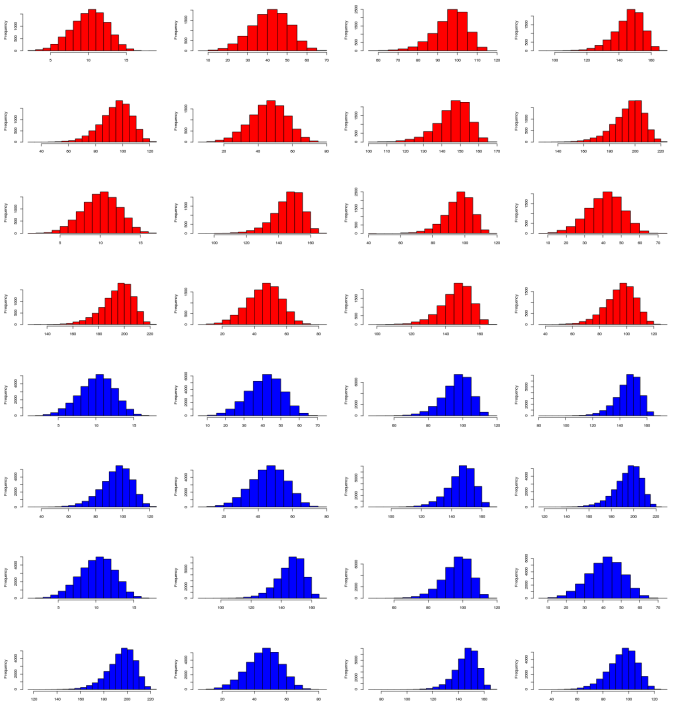

Figure 5. Random numbers generated for Weibull distribution four sets of training and test samples (lines) for four classes (columns).

The Weibull distribution is an important tool in the analysis of reliability and durability of equipment, such as resistance to fracture of the glass and flaws in semiconductors and capacitors. The histograms of random numbers generated for simulations using Continuous Uniform distribution are presented in Figure 5. In a simulation with one dimensional data and continuous uniform distribution, the best Kappa result was 66.72% with variance of 2.67×10^{-6}. The number of misclassifications was 29952. When dimension of data was increased for 2, the Kappa coefficient resulted in 99.93% with variance of 6.78×10^{-9}, and 55 misclassifications. For three dimensional data, the results were 99.48% and 5.74×10^{-8} for Kappa and its variance, respectively. For last, the Kappa coefficiente pointed out 99.98% with variance of 1.48×10^{-9} for 4 dimensional data.

TABLE VII. SUMMARY OF BEST RESULTS, BY STATISTICAL DISTRIBUTION, ACCORDING TO THE KAPPA COEFFICIENT

Statistical Distribution	Number of Dimensions	Kappa Coefficient
Negative Binomial	2 or more	$>$ 90.0 %
Logistic	2 or more	$>$ 95.0 %
Log-Normal	2 or more	$>$ 80.0 %
Gamma	3 or more	$>$ 95.0 %
Weibull	2 or more	$>$ 99.9 %

Table VII presents a summary of the best results obtained by each statistical distribution in the simulations. For the Negative Binomial and Logistic distributions used in the FGNB classifier with three or more dimensions, was possible to achieve more than 99% of correct classification, according to the Kappa Coefficient. In a similar way, using the Gamma distribution with four dimensions, the FGNB classifier achieve more than 99% of accuracy. However, for Log-Normal distribution, the FGNB performance is reasonable, but its results are better in higher dimensions of data. The classification of the Weibull distribution presented excellent results. From the dimension two, the Kappa coefficient obtained was above 99.9%.

V. CONCLUSION AND FUTURE WORK

In this paper, we presented an assessment of FGNB accuracy for classification tasks using data with different statistical distributions. We made simulations with five different statistical distributions: Negative Binomial, Logistic, Log-Normal, Weibull and Gamma. For each statistical distribution were analysed four different dimensions according to number of misclassifications, Kappa Coefficient and its variance. According to the results obtained, FGNB could be recommended to classify data from all distributions studied **in this paper**. For the distributions Negative Binomial, Logistic, Weibull and Gamma, the Kappa coefficient exceeded 90%. For the Log-Normal distribution, with maximum possible dimensions, the Kappa coefficient reached approximately 85%.

As future work, we would like to analyse the performance of this method with a database where each dimension can be given by a different statistical distribution. For instance, in a case of three dimensions, the first one would be a sample of Weibull distribution, the second one would be a sample of Gamma distribution and the last one would be a sample of Logistic distribution.

 ISBN: 978-1-61208-393-3

VI. ACKNOWLEDGMENTS

This project is partially supported by grants 310561/2012-4 and 310470/2012-9 of the National Council for Scientific and Technological Development (CNPq) and is related to the National Institute of Science and Technology "Medicine Assisted by Scientific Computing"(181813/2010-6) also supported by CNPq.

REFERENCES

[1] A. R. Webb and K. D. Copsey, Statistical Pattern Recognition., 3rd ed. Chichester: Wiley, 2011.

[2] R. O. Duda, P. E. Hart, and D. G.Stork, Pattern Classification, 2nd ed. Wiley Interscience, 2000.

[3] P. Domingos and M. Pazzani, "On the optimality of the simple bayesian classifier under zero-one loss." Machine Learning, no. 29, 1997, pp. 103–130.

[4] G. John and P. Langley, "Estimating continuous distributionsin bayesian classifiers." in 11th Conference on Uncertainy in Artificial Intelligence, Montreal, Canada, 1995, pp. 338–345.

[5] J. Hilden, "Statistical diagnosis based on conditional independence does not require it." Computers in Biology and Medicine, 14, 429–435, 1984.

[6] N. Friedman, D. Geiger, and M. Goldszmidt, "Bayesian network classifiers," Machine Learning, no. 29, 1997, pp. 131–163.

[7] S. Monti and G. F. Cooper, "A bayesian network classifier that combines a finite mixture model and a naive bayes in model." in 15th Conference on Uncertainy in Artificial Intelligence,, Stockholm, Sweden, 1999.

[8] C. M. Bishop, Pattern Recognition and Machine Learning., 1st ed. Singapure: Springer, 2007.

[9] R. M. Moraes and L. S. Machado, "Gaussian naive bayes for online training assessment in virtual reality-based simulators." Mathware & Soft Computing, no. 16, 2009, pp. 123–132.

[10] L. A. Zadeh, "Fuzzy sets," Information Control, no. 8, 1965, pp. 338–353.

[11] L. A. Zadeh., "Probability measures of fuzzy events," J. Math. Anal. Applic., no. 10, 1968, pp. 421–427.

[12] R. M. Moraes and L. S. Machado, "Fuzzy gaussian naive bayes applied to online assessment in virtual reality simulators." World Scientific, 2010, pp. 243–248.

[13] R. M. Moraes and L. S. Machado., "Online assessment in medical simulators based on virtual reality using fuzzy gaussian naive bayes," Journal of Multiple-Valued Logic and Soft Computing, no. 18, 2012, pp. 479–492.

[14] R. M. Moraes, "Performance analysis of evolving fuzzy neural networks for pattern recognition," Mathware & Soft Computing, no. 20, 2013, pp. 63–69.

[15] J. A. Ferreira, E. Soares, and R. M. Moraes, "Assessment of fuzzy gaussian naive bayes classifier using data with different statistical distributions," in III Congresso Brasileiro de Sistemas Fuzzy (CBSF2014), Joao Pessoa, Brazil, August 2014.

[16] M. Ramonia and P. Sebastiani, "Robust bayes classifiers," Artificial Intelligence, vol. 125, January 2001, pp. 209–226.

[17] R. A. Johnson and D. W. Wichern, Applied Multivariate Statistical Analysis, 4th ed. Prentice Hall, 1998.

[18] G. J. Klir and B. Yuan, Fuzzy Sets and Fuzzy Logic: Theory and Applications. Prentice Hall, 1995.

[19] J. Gentle, Elements of Computational Statistics. Springer, 2005.

[20] R Development Core Team. R: A language and environment for statistical computing., R Foundation for Statistical Computing, 2009.

[21] J. Cohen, "A coefficient of agreement for nominal scales," Educational and Psychological, no. 20, 1960, pp. 37–46.

[22] R. M. Moraes and L. S. Machado, "Psychomotor skills assessment in medical training based on virtual reality using a weighted possibilistic approach," Knowledge-Based Systems, no. 70, 2014, pp. 97–102.

[23] J. R. Landis and G. G. Koch, "The measurement of observer agreement for categorical data." Biometrics, vol. 33, no. 1, 1977, pp. 159–174.

 ISBN: 978-1-61208-393-3