Ciencia Latina
Revista Multidisciplinar

DOI: https://doi.org/10.37811/cl_rcm.v6i5.3159

Method of bayesian concordance and its application in problems of multiclass classification with unbalanced categories

Ricardo Borja-Robalino
ricardostalinborjar@gmail.com
https://orcid.org/0000-0002-3899-1140

Antonio Monleón-Getino
amonleong@ub.edu
https://orcid.org/0000-0001-8214-3205

José Rodellar
jose.rodellar@upc.edu
https://orcid.org/0000-0002-1514-7713

ABSTRACT

The Machine or Deep Learning classification techniques use several performance evaluations measures. The kappa index is a highly undervalued measure regardless of its reliability in problems with unbalanced classes. On the other hand, Bayesian methods generate great contributions to statistics, adding uncertainty to the probabilistic model that allows estimating parameters with better adjustments. This research offers an innovative alternative for researchers by designing a free access library in the RStudio environment that evaluates classifiers through a measure of Bayesian-frequentist agreement. It uses three Bayesian models (Dirichlet, Multinomial-Dirichlet and Beta) with the Markov Monte Carlo chain method. The library was applied to the classification of leukemic cells at the Hospital Clínic (Barcelona), demonstrating its effectiveness in using the Bayesian kappa index for unbalanced data in relation to other measures, as well as the robustness and sensitivity of the design. For teaching use, the library has an additional function that simulates classifiers through a multinomial distribution, allowing them to be evaluated.

Keywords: Markov chains Monte Carlo; Bayesian inference; kappa index; Bayes Theorem; Decision Theory.

Correspondencia: ricardostalinborjar@gmail.com
Artículo recibido 10 agosto 2022 Aceptado para publicación: 10 septiembre 2022
Conflictos de Interés: Ninguna que declarar

Cómo citar: Borja Robalino, R., Monleón Getino, A., & Rodellar, J. (2022). Método de concordancia bayesiana y su aplicación en problemas de clasificación multiclase con categorías desequilibradas. Ciencia Latina Revista Científica Multidisciplinar, 6(5), 1064-1090. https://doi.org/10.37811/cl_rcm.v6i5.3159

Método de concordancia bayesiana y su aplicación en problemas de clasificación multiclase con categorías desequilibradas

RESUMEN

Las técnicas de clasificación Machine o Deep Learning utilizan varias medidas de evaluación del rendimiento. La índice kappa es una medida muy infravalorada independientemente de su fiabilidad en problemas con clases desequilibradas. Por otro lado, los métodos bayesianos generan grandes aportes a la estadística, agregando incertidumbre al modelo probabilístico que permite estimar parámetros con mejores ajustes. Esta investigación ofrece una alternativa innovadora para los investigadores al diseñar una biblioteca de libre acceso en el entorno RStudio que evalúa clasificadores a través de una medida de concordancia bayesiana-frecuentista. Utiliza tres modelos Bayesianos (Dirichlet, Multinomial-Dirichlet y Beta) con el método de cadena Markov Monte Carlo. La biblioteca se aplicó a la clasificación de células leucémicas en el Hospital Clínic (Barcelona), demostrando su eficacia en el uso del índice bayesiano kappa para datos desequilibrados en relación con otras medidas, así como la robustez y sensibilidad del diseño. Para uso docente, la biblioteca cuenta con una función adicional que simula clasificadores a través de una distribución multinomial, lo que permite evaluarlos.

Palabras clave: *cadenas de markov monte carlo; inferencia bayesiana; indice kappa; teorema de bayes; teoría de la decisión.*

1. INTRODUCTION

Currently, multiclass classification problems are essentially focused on the development and continuous improvement of machine learning algorithms applied to large volumes of data (Maxwell et al., 2018). However, the efficiency of these classifiers is affected by the continuous occurrence of cases with categories significantly less represented than others (unbalanced). This has promoted techniques that reduce the distance in relation to proportions between classes, such as the algorithm Adaboost and others. In the same way, these alternative solutions still generate a crossroad for the research with very few options when evaluating classifiers (Shuo & Xin, 2012). Making accuracy the most widely used discriminatory measure when comparing observers or classification algorithms. In conceptual terms, the accuracy represent an overall measure of how close the result is with respect to a reference (Westgard, 2008).

An alternative is the Cohen's proposal with the kappa index as a measure of the concordance analysis between two human observers or classifiers. It offers a reliable comparison for the unbalanced multiclass case to know which is the best or worst classifier through the agreement observed, corrected by random effects that present susceptibility to biased classes. It is applicable to all areas, especially in medical areas such as the case of diagnosis and interpretation of findings related to examinations. Kappa formulation relates to the agreement observed as the number of elements correctly classified and what would be expected by chance to the instances of each class together with the elements that the observer or classifier agreed with the absolute truth (McHugh, 2012). The kappa index is taken as a parameter of reliability (accuracy) in cases where the absolute truth is known (gold standard). Otherwise, its validity is treated under the sensitivity and specificity (Brennan & Prediger, 1981).

Based on frequentist theory, the kappa index uses the information of a sample, based on the probability that an event occurs in relation to the pattern observed so far. However, the appearance of the Bayesian method, which introduces information such as the degree of belief that is given to an event through previous knowledge, expectations, experience of the researcher and others, make this alternative approach highly promising.

The Bayesian inference assumes the parameter θ as a random variable under a certain distribution $f(\theta)$, computed through the product of a prior distribution $P(\theta)$ by the probabilistic model (Likelihood), which describes the prior knowledge of the variable of interest contained in the a posteriori distribution $\pi(\theta|x)$. The estimation of the desired value is a decision problem that uses the Bayes Theorem, allowing to achieve solid and robust results.

This paper focuses on the analysis of agreement for the case of unbalanced classes using Cohen's kappa index, using Bayesian methods, showing robustness and effectiveness when making specific calculations according to observers or classifiers. The specific objective is the design of the KFreqBay library in RStudio, which allows concordance analysis (kappa index), applicable to multiclass data with unbalanced categories. The library includes both the frequentist and Bayesian method with Markov Monte Carlo chains (MCMC), which can simulate a gold standard and classifiers with multinomial distribution or work with a set of data preset by the user. Three Bayesian models are applied: two Dirichlet distributions, one Dirichlet mixture with one Multinomial, and finally the use of two Beta distributions. They allow to effectively estimate the kappa index.

It is important to emphasize that the library developed is unique at the level of the R software, considering that it merges the two methods (frequentist and Bayesian) in the evaluation of classifiers. The user does not need the tedious job of installing several additional libraries. The library presents an additional report and relevant images in pdf format with the basic and necessary statistics of each pair of existing classifiers in the database, thus reducing working time of researchers when issuing a decision based on the best or worst observer or classifier algorithm.

The library was validated in two ways: 1) applied to a set of values obtained by simulation; and 2) applied to the classification results of an unbalanced database of digital microscope images of leukemic cells from peripheral blood of patients of the Hospital Clinic (Barcelona - Spain). Classifications were performed using machine learning techniques such as Linear Discriminant Analysis, Support Vector Machine and Random Forest.

2. THEORETICAL BACKGROUND

2.1 Concordance in categorical data

The concordance can be defined as the analysis that allows to measure the degree of agreement between two or more classifiers or observers, determining to what extent their results coincide in relation to the same phenomenon (Fleiss et al., 2003).

The Cohen kappa index is used in dichotomous cases and constitutes the observed agreement corrected for the effects of chance, proposing a standardized measure between [-1, 1]. It is formulated from the contingency table or well-known as the confusion matrix, which represents the frequency of hits and disagreements between methods for each category analyzed (See Table 1).

Table 1. *Table of frequencies - dichotomous variables*

		OBSERVER 1		
		POSITIVE	NEGATIVE	TOTAL
OBSERVER 2	POSITIVE	f_{11}	f_{10}	F_{1T}
	NEGATIVE	f_{01}	f_{00}	F_{0T}
	TOTAL	f_{1T}	f_{0T}	N

The most common agreement measures are (Borja, 2019):

- Index Kappa: $$k = \frac{Po-Pe}{1-Pe}$$

where:

P_o = Proportion of observed agreement.

$$Po = \frac{f_{11} + f_{00}}{N}$$

P_e = Proportion of expected agreement by chance (product of marginal frequencies).

$$Pe = \left(\frac{F_{1T}}{N}\right) * \left(\frac{f_{1T}}{N}\right) + \left(\frac{F_{0T}}{N}\right) * \left(\frac{f_{0T}}{N}\right) = \frac{F_{1T*}f_{1T} + F_{0T*}f_{0T}}{N^2}$$

- Classification error or average error. - Proportion of misclassified cases.

$$\frac{f_{10+}f_{01}}{N}$$

- Positive True or sensitive. - Proportion of positive cases well classified.

$$\frac{f_{11}}{f_{11} + f_{01}}$$

- Negative True or specificity. - Proportion of well-ranked negative cases.

$$\frac{f_{00}}{f_{00}+f_{10}}$$

- False Positive. - Proportion of misclassified positive cases (Type I error).

$$\frac{f_{10}}{f_{10}+f_{00}} = 1 - specificity$$

- False Negative. - Proportion of badly classified negative cases (Type II error).

$$\frac{f_{01}}{f_{01}+f_{11}}$$

- Accuracy. - It indicates the degree of reproducibility of responses between observers.

$$\frac{f_{11}+f_{00}}{N}$$

- Confidence interval of the kappa coefficient (95%). - Corresponds to kappa ± the approximate standard error.

$$CI = k \ \pm z_{1-\alpha/2}\sqrt{\frac{Po(1-Po)}{N(1-Pe)^2}}$$

- McNemar test. - Determines whether or not there is a systematic difference between two observers.

$$z^2 = \frac{d^2}{Var(d)} = \frac{(f_{10}-f_{01})^2}{f_{10}+f_{01}}$$

The value of the agreement k increases while the classes are distributed asymmetrically by the observers. Its effect is contrary to the measure that increases the number of classes and even more if they are biased, showing great sensitivity to unbalanced cases (Watson & Petrie, 2010). For the assessment of the degree of agreement, the proposal of Landis & Koch (1977) is used, assuming that the agreement is exactly what was expected by chance in the case of having $k = 0$ (see Table 2).

Table 2. *Valuation table of kappa index*

Kappa	Degree of agreement
< 0	Without agreement (less than expected by chance)
(0 – 0.2]	Insignificant.
(0.2 – 0.4]	Low
(0.4 – 0.6]	Moderate
(0.6 – 0.8]	Good
(0.8 – 1]	Very good

In cases of comparison of more than two evaluators (multinomial), the main measure of agreement is the Fleiss kappa index. It represents the corrected observed agreement between classifiers in the case where all the evaluators take a random result (Garabedian et al., 2017). It is formulated as:

$$k = 1 - \frac{nm^2 - \sum_{i=1}^{n}\sum_{j=1}^{r} x_{ij}^2}{nm(m-1)\sum_{j=1}^{r} \bar{p}_j \bar{q}_j}$$

where:

- *r*= number of categories; *p* = proportion of positive agreements; q= proportion of negative agreements.
- *n*= number of samples; *m* = the number of trials of each evaluator for each case.
- x_{ij} = the number of observers who assign the i-th subject to the j-th category.

A hypothetical example in the dichotomous case may be the need to know if a new image processing equipment, which allows detecting lung cancer more economically and quickly, can replace the old device (gold standard). For this, 900 images have been analyzed with the two teams, obtaining the following results (see Table 3):

Table 3. *Results of the illustrative example*

		EQUIPMENT 1		
		POSITIVE	NEGATIVE	TOTAL
EQUIPMENT 2	POSITIVE	59	12	**71**
	NEGATIVE	4	825	**829**
	TOTAL	**63**	**837**	**900**

Applying the equations (2.1, 2.5, 2.6, 2.7, 2.8, 2.9 and 2.10) we have:

Kappa: $k = \frac{0.982-0.8621}{1-0.8621}$ =0.8712 ; Confidence interval: $CI = [0.841,0927]$

Sensitivity: $\frac{59}{59+4}$ =0.93 ; Specificity: $\frac{825}{825+12} = 0.98$; Accuracy: $\frac{59+825}{900}$ =0.982

The results show that the new equipment has a high accuracy. However, taking into account that we work with unbalanced classes, we observe the kappa index where we can assess as a very good agreement between two devices (87.12%), correcting the effect of chance (86.21%). It indicates that there is greater accuracy when the team gives negative the existence of lung cancer (98%), more than in a positive response (93%). This concludes that it is an excellent option to replace the old equipment.

2.2 Probability distributions

This section presents a summary of the main probability distributions related to categorical variables in the Bayesian environment.

Multinomial Distribution

The multinomial distribution (Sheldom, 2014) is a generalization of the binomial for a multinomial random variable $= x_{1,}x_{2,\ldots.}x_k$, with k excluding events $S_1, S_2 \ldots . S_k$, respective probabilities $p_1, p_2 \ldots . p_k$:

$P(S_1) = p_1$, $P(S_2) = p_2 \ldots . . P(S_k) = p_k$, $\therefore$, $\sum_{i=1}^{k} P(S_i) = 1$

The probability that the event $S_1 \ldots . S_k$ happens $R_1 \ldots \ldots R_k$ times, successively forming a partition of the sample space Ω, is called multinomial distribution and its mass function is:

$$f(x_1, x_2 \ldots x_k) = P[(S_1 = R_1) \cap (S_1 = R_2) \cap \ldots . \cap (S_k = R_k)]$$

$$= \frac{n!}{x_1! \, x_2! \ldots . . x_k!} p_1^{x_1} p_2^{x_2} \ldots . p_k^{x_k}$$

where:

- $\sum_{i=1}^{k} R_i = n$
- For k=2 it is reduced to a binomial distribution.

Beta Distribution

It is widely used in continuous variables with restrictions in a range of length (0.1), and the most used in Bayesian inference as a priori distribution, due to its good adjustment to a wide variety of empirical distributions (Gupta & Nadarajah, 2004).

In the beta distribution $X \sim Beta(\alpha, \beta)$its density function is:

$$f(x) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\,\Gamma(\beta)}\, x^{\alpha-1}(1-x)^{\beta-1}$$

where:

- Γ is the gamma function.
- $0 > x < 1 \quad ; \qquad \alpha, \beta > 0$.
- α y β are profile parameters.
- It is asymmetric if $= \beta$, with $\alpha < \beta$ have positive asymmetry and $\alpha > \beta$ negative asymmetry.
- If $\alpha = \beta = 1$ then $X \sim U(0,1)$.

Dirichlet distribution

This distribution is one of the most used within Bayesian inference as a priori distribution representing uncertainty in results of categorical and multinomial distributions (Blei, Nigle, & Jordan, 2003). It is the multivariate generalization of the beta distribution (k = 2) and of a continuous multivariate family. Its density function is:

$$f(X,\alpha) = \frac{1}{Beta(\alpha)} \prod_{i=1}^{k} x_i^{\alpha_i - 1} \qquad Beta(\alpha) = \frac{\prod_{i=1}^{k} \Gamma(\alpha_i)}{\Gamma(\sum_{i=1}^{k} \alpha} \quad (2.15)$$

$$f(X,\alpha) = \frac{\Gamma(\sum_{i=1}^{k} \alpha_i)}{\prod_{i=1}^{k} \Gamma(\alpha_i)} \prod_{i=1}^{k} x_i^{\alpha_i - 1}$$

where:

- $x_1, x_2, x_3 \ldots . x_k > 0$; $\sum_{i=1}^{k} x_i = 1$ for all $i \, \epsilon \, [1, k]$
- Probability of each category= $\frac{\alpha_i}{\sum_{i=1}^{k} \alpha_i}$; $k =$ number of categories.
- $\alpha_1, \alpha_2, \alpha_3 \ldots . \alpha_k > 0$

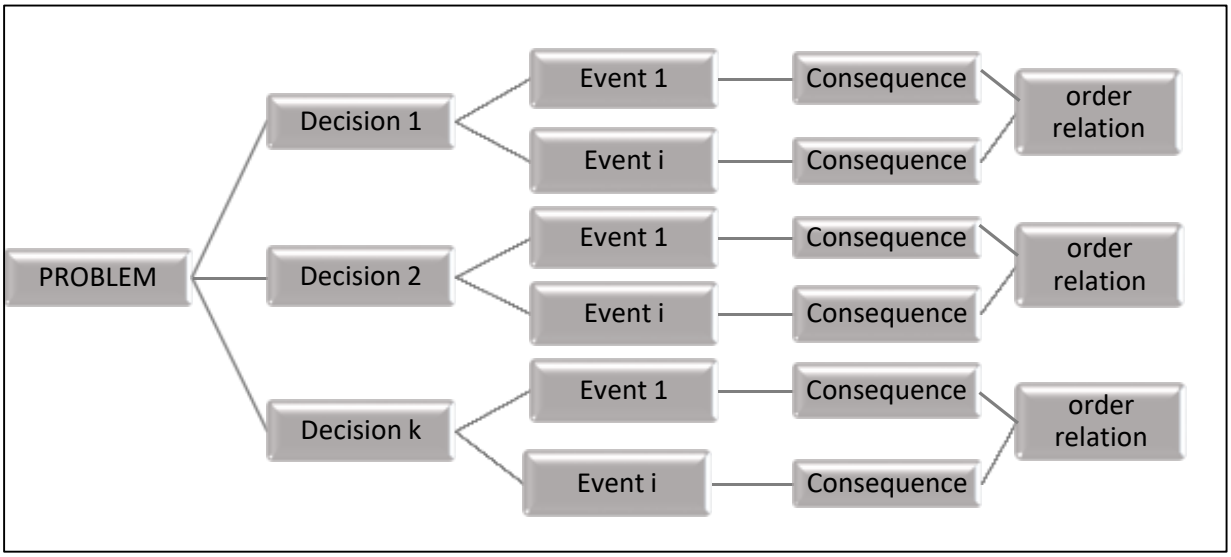
2.3 Bayesian inference

It is a statistical method that allows obtaining a more precise prediction of a parameter θ of interest (posteriori), adding previous information of the event (priori) to the probabilistic model (likelihood). The Bayesian inference is characterized by assuming θ as a random variable under a certain distribution $f(\theta)$. The estimation of the desired value is a decision problem that uses the Bayes Theorem, allowing to achieve solid and robust results. This type of inference allows introduce uncertainty into the data and regulate predictions (Shridhar et al., 2019), adjusting the parameters of the distribution in the continuous case or by depending on the prevalence of the classes in the categorical case. It is very useful for unbalanced multiclass cases (Sanjib et al., 2000).

2.3.1 Decision theory

It is a process based on established criteria that allows responding with the highest reliability to an observer who faces a decision problem in an environment of uncertainty. It is defined as a quatrain $(\mathbb{D}, \mathbb{E}, \mathbb{C}, \succ)$, which starts with a problem that comprises a set of decisions$\mathbb{D} = \{D1, D2 \dots Dk\}$, associated with a set of relevant uncertain events $E_k = \{E_{k1}, E_{k2} \dots \dots E_{ki}\} \subset \mathbb{E}$. Each event has a consequence $\mathbb{C}$, which if there is more than one the order relation is used $\succ$, which determines which is the most appropriate (Laurence & Pascal, 2009) (see Figure 1).

Figure 1: Decision tree



2.3.2 Bayes theorem

Fulfilling the assumptions of disjoint and exhaustive events, Bayes proposes the following theorem that presents the probability of a random event $X = x_1, x_2, \ldots . x_n$ mutually exclusive given Y in terms of the conditional probability of the event Y given X and the marginal distribution of X (Bradley 2013; Press, 2009):

$$P(x_k|Y) = \frac{P(Y|x_k)P(x_k)}{\sum_{i=1}^{k} P(x_i)P(Y|x_i)} \quad ; k = 1,2..n$$

where $P(x_k|Y)$ = a posteriori probability, $P(Y|x_k)$ = conditional probability, $P(x_k)$= a priori probability and $\sum_{i=1}^{k} P(x_i)P(Y|x_i)$= total probability.

In a simpler and concise way, it can be stated:

$$A\ posteriori\ probability\ \propto Likelihood * a\ priori\ probability$$

The assigned a priori distribution can be of three types: informative if it incorporates information from previous analyzes, not informative if it is constructed based on subjective considerations and finally of structural type in the case that it incorporates information on relationships between parameters (D`Agostini, 2003).

The calculation of a posteriori probability starting from a priori generates multiple numerical difficulties that can trigger illogical results and with great complexity in their interpretation. However, this shortcoming can be covered by working with conjugated distributions that comply with the following property:

A family $\mathcal{P}$ of a prior distributions on X is said to be conjugated for sampling if: for any prior in $\mathcal{P}$, the corresponding posteriori also belongs to $\mathcal{P}$ (Cristóbal, 2000).

The Table 4 presents different mixtures of conjugated families.

Table 4. *Conjugated families (Cristóbal, 2000).*

Priori	Likelihood	Posteriori	Non-informative Prior parameter
Beta	Bernoulli	Beta	$\alpha = \frac{1}{2}, \beta = \frac{1}{2}$
Dirichlet	Multinomial	Dirichlet	$\alpha_i = \frac{1}{2}; (i = 1 \dots m)$
Multinomial	Dirichlet	Beta	$\alpha = \frac{1}{2}, \beta = \frac{1}{2}$
Gamma	Poisson	Gamma	$a = 0; \quad p = \frac{1}{2}$
Normal	Normal	Normal	$\sigma_0 \to \infty$
Gamma	Normal	Gamma	$a = 0; \quad p = 0$
Beta	Binomial	Beta	$\alpha = \frac{1}{2}, \beta = \frac{1}{2}$
Binomial	Binomial	Beta	$\alpha = \frac{1}{2}, \beta = \frac{1}{2}$

2.3.3 Markov Chains Monte Carlo

Markov Chains describe a discrete stochastic process that evolves probabilistically over time (Hillier & Lieberman, 2010), where, the probability of a subsequent event x_{n+1} depends on the immediately preceding event x_n (markovian property). Generating a short memory effect in the chains that allow conditioning future probabilities:

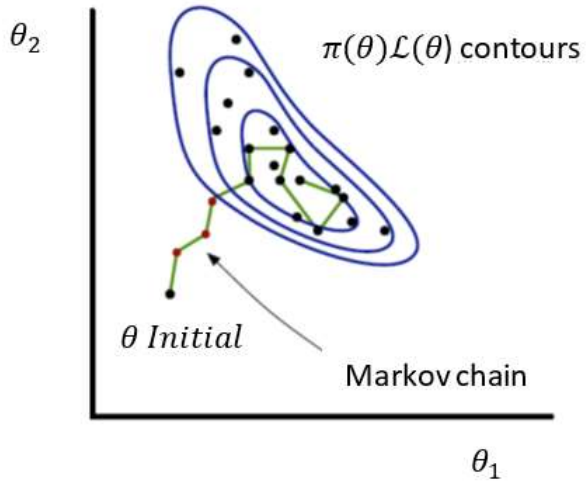
$$P(x_{n+1} = x_{n+1} | x_n = x_n)$$

Monte Carlo simulation is defined as the way to estimate a fixed parameter through the repeated generation of random numbers (Chib et al., 2002).

Monte Carlo Markov Chains (MCMC) are defined as a simulation method that allows generate samples of the distribution afterwards, estimating quantities or parameters of interest through random sampling in a probabilistic space. MCMC are used in Bayesian inference to solve the difficult task of calculating the a posteriori probability of the Bayes Theorem, in cases with complex distributions. MCMC perform a series of repetitions of n points of the M-dimensional space through a random number generator, recognizing the behavior of the system (Lebreton et al., 2004). Calculations can be developed through several algorithms, the most common being Gibbs sampling, which is considered as a particular case of the Hasting Metropolis.

The algorithm has a burn-in phase, which is the process that accelerates the convergence of the chain by eliminating points that are outside the contour of the stationary process, due to its low probability when starting the algorithm. For the diagnosis of convergence of one or more Markov chains to the estimated value, Gelman Rubin scale reduction factors are commonly used to compare variations within and between the chains. Figure 2 represents a two-dimensional MCMC algorithm.

Figure 2. *Two-dimensional MCMC algorithm (Ford, 2015).*



3. DEVELOPMENT OF THE LIBRARY

The motivation for the library is to help solve common real-life problems in relation to multiclass classification with unbalanced categories due to the continuous development of new machine learning algorithms. Focused on the methods of concordance with the application of statistical inference and the punctual estimation of the kappa index and other general statistics, the library allows a robust and efficient concordance analysis. It is applicable to data with dichotomous and politomic variables with unbalanced categories using the Frequentist and Bayesian method using Monte Carlo Markov chains (MCMC), either by creating a standard gold and classifiers with multinomial distribution by simulation or through a set of preset data. The KfreqBay library has a wide range of use either for research, educational or other applications, in general related to the evaluation of unbalanced multiclass classifiers based on concordance analysis, allowing both frequent and Bayesian perspectives in a robust, efficient way and fast.

In practice, Bayesian inference was implemented to estimate the Cohen's kappa index, by designing a library in the R language, obtaining as a result a frequentist and Bayesian concordance analysis, very effective in the unbalanced multiclass case.

The software JAGS (Just Another Gibbs Sampler) linked to the integrated development environment RStudio with the rjags library, made it possible to analyze Bayesian hierarchical models by applying Monte Carlo Markov Chains using the Gibbs sampling algorithm. We worked with the Cohen kappa index comparing several classifiers in pairs, because Fleiss multinomial kappa can sometimes return low values even when the agreement is really high (Powers, 2012).

A number of primary and secondary functions were programmed that, in the first instance, convert the input data into the appropriate format for the respective calculation. The frequentist analysis was made by extracting the general statistics, and a descriptive graphical analysis of the proportion of the classes was obtained. Three Bayesian models were developed that estimate the parameter of interest, demonstrating robustness and sensitivity of the proposed model with the significant contribution of the chosen distributions for the likelihood. The library allows to add information about the prevalence of the classes in the form of probability at the moment of performing the Bayesian calculation.

For educational and experimental purposes, we have the option of simulating the response of a classifier through a multinomial distribution, building the gold standard and several observers according to the characteristics pre-established by the user. Consequently, the analysis of frequentist and Bayesian concordance is carried out, simultaneously.

3.1 Bayesian models

In order to achieve logical and interpretable results, the Bayesian models were based on the mixture of conjugated families in Table 4 proposed by Cristóbal (2000). It was designed in text format using the function **textconnection (model)** for its analysis within the JAGS environment.

For the three models, the following likelihood function was proposed:

$$\mathcal{L} = \left(\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\,\Gamma(\beta)}\,x^{\alpha-1}(1-x)^{\beta-1}\right)^{x}\left(1-\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\,\Gamma(\beta)}\,x^{\alpha-1}(1-x)^{\beta-1}\right)^{1-x}$$

This is formulated starting from a Bernoulli distribution $(p) = p^x(1-p)^{1-x}$, with $p{\sim}Beta(1,1) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\,\Gamma(\beta)}\,x^{\alpha-1}(1-x)^{\beta-1} \approx Uni\,(0,1)$, considering that we work with categorical variables in the dichotomous case and with values generally between (0,1).

For the first model we used a Dirichlet distribution:

$$f(X,\alpha) = \frac{\Gamma(\sum_{i=1}^{k}\alpha_i)}{\prod_{i=1}^{k}\Gamma(\alpha_i)}\prod_{i=1}^{k}x_i^{\alpha_i-1}$$

Applying the Bayes theorem we may write:

$$\pi(X|Y) = \left(\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\,\Gamma(\beta)}\,x^{\alpha-1}(1-x)^{\beta-1}\right)^{x}\left(1 - \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\,\Gamma(\beta)}\,x^{\alpha-1}(1-x)^{\beta-1}\right)^{1-x}\frac{\Gamma(\sum_{i=1}^{k}\alpha_i)}{\prod_{i=1}^{k}\Gamma(\alpha_i)}\prod_{i=1}^{k}x_i^{\alpha_i-1}$$

Regarding the programming of the models, only the first will be detailed. Using equation (2.1) of the kappa index, the following algorithm was proposed:

```
#Programming in R
Model <- "model {
    # Verosimilitud
    kappa <- (p_agreement - expected_agreement) / (1 -expected_agreement)
    expected_agreement <- sum(p1 * p2)
    for (i in 1:n_ratings) {
    rater1[i] ~ dcat(p1)
    rater2[i] ~ dcat(p2)
    agreement[i] ~ dbern(p_agreement) }
    #  Parámetros priori
    p1 ~ ddirch(alpha)
    p2 ~ ddirch(alpha)
    p_agreement ~ dbeta(1, 1)
    alpha <- prob }"
```

For the second model a function is proposed that represents the mixture of a Dirichlet - Multinomial distributions, described and developed by Monleón-Getino (2018) and Monleón-Getino et al. (2019):

$$p_i(x) = \frac{(\mathrm{N}!)\Gamma(\sum_i^k \alpha_i)}{\Gamma(\mathrm{n}+\sum_i^k \alpha_i)}\prod_i^k\left(\frac{\Gamma(\sum_i^k x_i+\alpha_i)}{(x_i!)\Gamma(\alpha_i)}\right)$$

Therefore, by using the Bayes theorem, we have:

$$\pi(X|Y) = \prod_{i=1}^{n}\left(\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\,\Gamma(\beta)}\,x_i^{\alpha-1}(1-x_i)^{\beta-1}\right)^{x_i}\left(1 - \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\,\Gamma(\beta)}\,x_i^{\alpha-1}(1-x_i)^{\beta-1}\right)^{1-x_i} * \frac{(\mathrm{N}!)\Gamma(\sum_i^k \alpha_i)}{\Gamma(\mathrm{n}+\sum_i^k \alpha_i)}\prod_i^k\left(\frac{\Gamma(\sum_i^k x_i+\alpha_i)}{(x_i!)\Gamma(\alpha_i)}\right)$$

For the third model, we worked with two prior Beta distributions with density function:

$$f(x) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\,\Gamma(\beta)}\,x^{\alpha-1}(1-x)^{\beta-1}$$

Then,

$$\pi(X|Y) = \left(\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\,\Gamma(\beta)}\,x^{\alpha-1}(1-x)^{\beta-1}\right)^{x}\left(1 - \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\,\Gamma(\beta)}\,x^{\alpha-1}(1-x)^{\beta-1}\right)^{1-x} * \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\,\Gamma(\beta)}\,x^{\alpha-1}(1-x)^{\beta-1}$$

Therefore, it was assumed that the responses of each classifier follow a previous distribution (beliefs), with prevalence introduced by the parameters of the priori. Together with our probabilistic model, they adjust the kappa estimate to different realities.

3.2 Library in the RStudio environment

The library created in R language has the name of KfreqBay, is free access, in zip format, installable in RStudio and downloadable from the address: https://github.com/RicardoBorja. It has the function K_Freq_Bay that allows running a frequentist and Bayesian concordance analysis with either a specific database or by simulating a gold standard and observers. It also includes a help menu (? K_Freq_Bay), which allows the user to know information about the parameters and illustrative examples that familiarize them with the process.

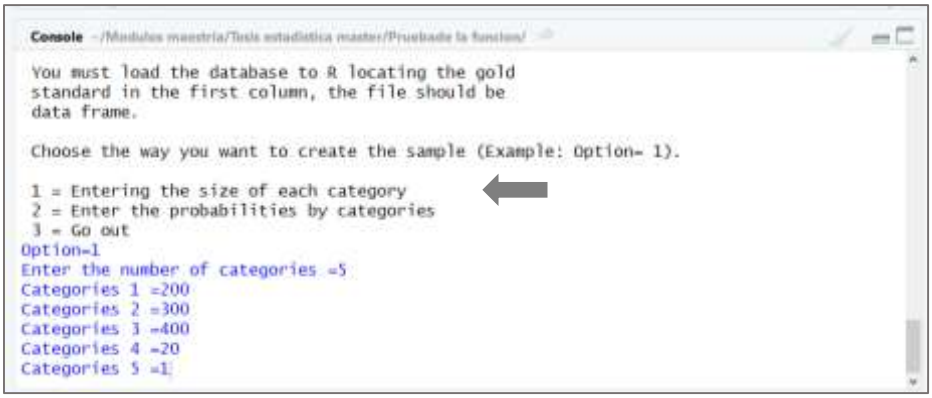
The K_Freq_Bay function has default values, with accessibility to changes according to the needs of the user. The function has the following form, where the arguments used are included:

```
K_Freq_Bay(data=FALSE,setseed=1234,num_mult=1000,burn_in=10000,chains=2,updat=
        1000,thin_=1,iter_thin_=20000,models=1, DIC_=0)
```

The designed library avoids the user the tedious activity of installing additional functions that a package requires for its proper functioning, automatically installing or activating everything necessary for the optimal execution of K_freq_Bay. In addition, it has a friendly environment that guides the process step by step, in the two cases of simulation or the use of a specific database. The required information is entered through the keyboard in numerical form.

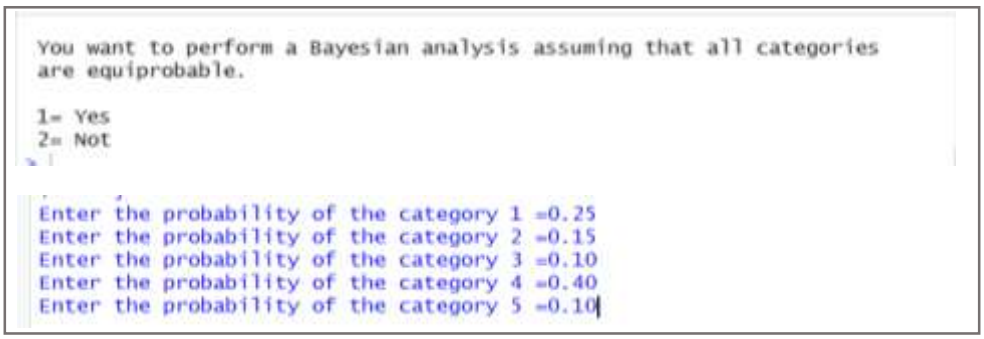
In the case of simulation of a gold standard and classifiers, the process starts with previous information, followed by two options to create the sample: enter the size of the categories or their probabilities (see Figure3).

Figure 3: *Generation of data through the number of categories and sample size*



Consequently, the number of observers and the desired precision are chosen. Once the database is created by simulation, the frequentist and Bayesian analysis is carried out, requesting if the user wishes that the prevalence of the a priori distribution is equiprobable or not. Figure 4 shows the case of information addition.

Figure 4. *Bayesian analysis including a prior information*



For the case of analyzing a given database, the previous information is known and the user goes directly to the option of Figure 4.

The library in any of the two cases presents as outputs a descriptive graph of the proportion of the classes, density graphs of the frequentist versus Bayesian kappa index, graphs of convergence diagnosis of Gelman Rubin, self-correlation, stationarity and final report of statistical values frequentist and Bayesian generals. They are generated for all possible pairs of classifiers (with the gold standard and with each other), with pdf format in the work folder. In addition, in the environment RStudio returns a list with:

1. Report of Gelman Rubin, Raftery Lewis and Cramer Von Mises (Methods to assessing Markov Chain Convergence).
2. Final report of the general statistics.
3. Final report in case of sample size changes.

In the case of simulation, at the end of the process the user has the option of changing the sample size while maintaining the same probabilities in the classes. This allows to know the different variations according to the increase or decrease of data.

A more detailed explanation is available in the thesis project through the link: https://upcommons.upc.edu/handle/2117/127344.

4. RESULTS

We evaluated the accuracy and sensitivity in the estimate of the Bayesian kappa index with the KFreqBay library, presenting different use cases. Three observers were simulated with a gold standard and five categories, under two scenarios: the first with frequencies of 200, 300, 400, 20, 1; in the second, the probability of each class was retained and the sample size was changed from 921 to 9000.

In the Bayesian part of each process, tests were carried out assuming equiprobability and with prevalence of 0.15, 0.40, 0.05, 0.20 and 0.20 in the classes for a prior distribution. Tables 5 and 6 summarize the results obtained in the estimation of the kappa index only of the gold standard compared to the first classifier. Final reports and graphs of all pairs of observers can be observed at: https://github.com/RicardoBorja.

Table 5. *Results frequentist method - validation process*

FREQUENTIST METHOD					
SAMPLE SIZE	PAIR OF OBSER.	KAPPA LOWER	KAPPA	KAPPA UPPER	ACCURACY
921	1-2	0.8622	0.8873	0.9124	0.9251
9000	1-2	0.8720	0.8802	0.8885	0.92

Table 6. *Bayesian method results with three models - validation process*

BAYESIAN METHOD							
SAMPLE SIZE	MODEL	PAIR OF OBSER.	KAPPA LOWER	KAPPA	KAPPA UPPER	P-VALUE 2 CHAINS	EQUIPROBABLE CATEGORIES
921	DI-DI	1-2	0.8591	0.8867	0.9106	(0.37;0.13)	YES
921	DI-DI	1-2	0.8590	0.8864	0.9104	(0.33;0.41)	NOT
921	DI-MUL	1-2	0.8030	0.9084	0.9339	(0.64;0.30)	YES
921	DI-MUL	1-2	-2.7271	0.9120	0.9372	(0.60;0.72)	NOT
921	BE-BE	1-2	0.9062	0.9244	0.9404	(0.27;0.05)	YES
921	BE-BE	1-2	0.9061	0.9245	0.9404	(0.43;0.96)	NOT
9000	DI-DI	1-2	0.8716	0.8802	0.8884	(0.95;0.89)	YES
9000	DI-DI	1-2	0.8718	0.8802	0.8884	(0.34;0.70)	NOT
9000	DI-MUL	1-2	0.7984	0.9047	0.9215	(0.20;0.39)	YES
9000	DI-MUL	1-2	-2.7848	0.9108	0.9239	(0.84;0.88)	NOT
9000	BE - BE	1-2	0.9142	0.9199	0.9254	(0.17;0.56)	YES
9000	BE-BE	1-2	0.9142	0.9200	0.9255	(0.11;0.09)	NOT

As observed in Tables 5 and 6, the reports generated by the three models were analyzed by checking the robustness and sensitivity of the proposed distributions within the probabilistic model and a prior probability in unbalanced multiclass cases. The Dirichlet-Dirichlet model presented a mesokurtic density with greater stability in both sample sizes, whereas the Dirichlet - Multinomial was leptokurtic in the equiprobable case and totally opposed when entering information. Finally, the Beta - Beta model presented a very narrow credibility interval that makes it very restrictive. In all cases, the chains converge to the estimated value with good precision in relation to the frequentist method. However, the Dirichlet - Dirichlet distribution is considered the most optimal and stable distribution for calculating the Bayesian kappa index.

4.1 Application of the Bayesian concordance analysis by K_Freq_Bay to the database of classification of leukemic cells in peripheral blood – Hospital Clinic.

The library was applied to the results of the automatic classification of peripheral blood digital images used for the initial diagnosis of leukemias and lymphomas (Boldú et al., 2019). They were obtained by the Cellsilab group formed by researchers from the CORE Laboratory of the Biomedical Diagnostic Center of the Hospital Clinic of Barcelona and the Mathematics Department of the Technical University of Catalonia.

The classifications were generated by three types of machine Learning algorithms (Linear Discriminant Analysis LDA, Support Vector Machine SVM and Random Forest RF) before and after the application of techniques of down – sampling and up – sampling to compensate for unbalanced classes. In this way six classification results were available for our study. For the gold standard we worked with 4365 data distributed in four categories: CLR reactive cells (338), acute lymphoid leukemia LAL (521), acute myeloid leukemia LAM (2839) and acute myeloid leukemia promyelocytic LAM_ PROM (667) (see Figure 5).

Figure 5. *Frequency chart - gold standard*

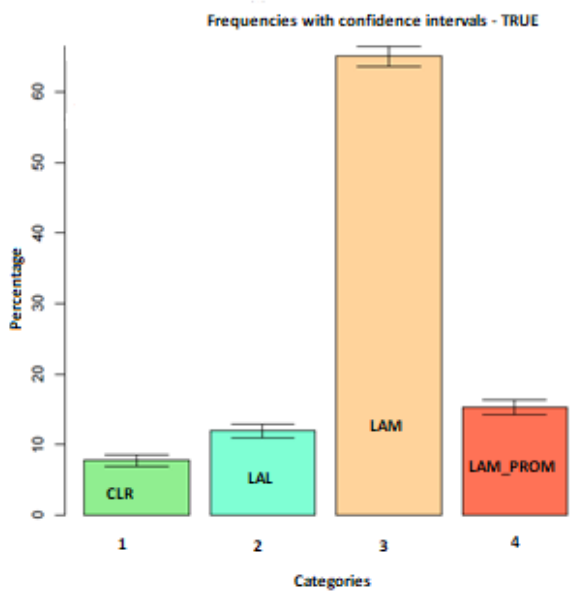
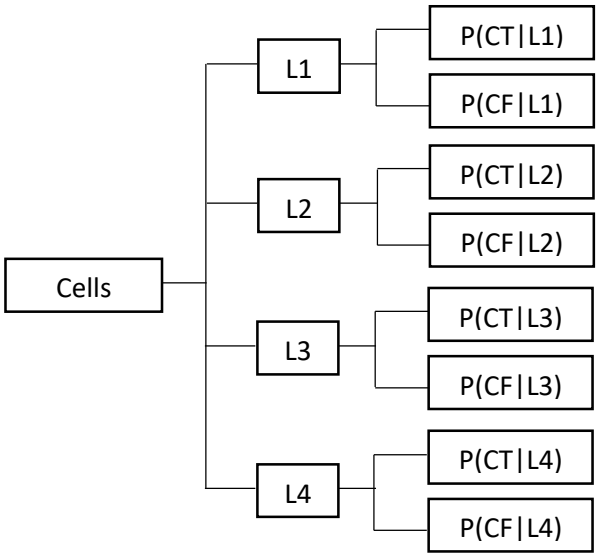


Figure 6 shows a decision tree that expresses the problem posed, taking the following considerations: CLR =L1, LAL= L2, LAM=L3 Y LAM_PROM = L4, CT = The classifier gave a positive interpretation of the cell when it was correct, CF = The classifier gave a negative interpretation of the cell when it was incorrect (see Figure 6).

Figure 6. *Decision tree – application to leukemic cells*



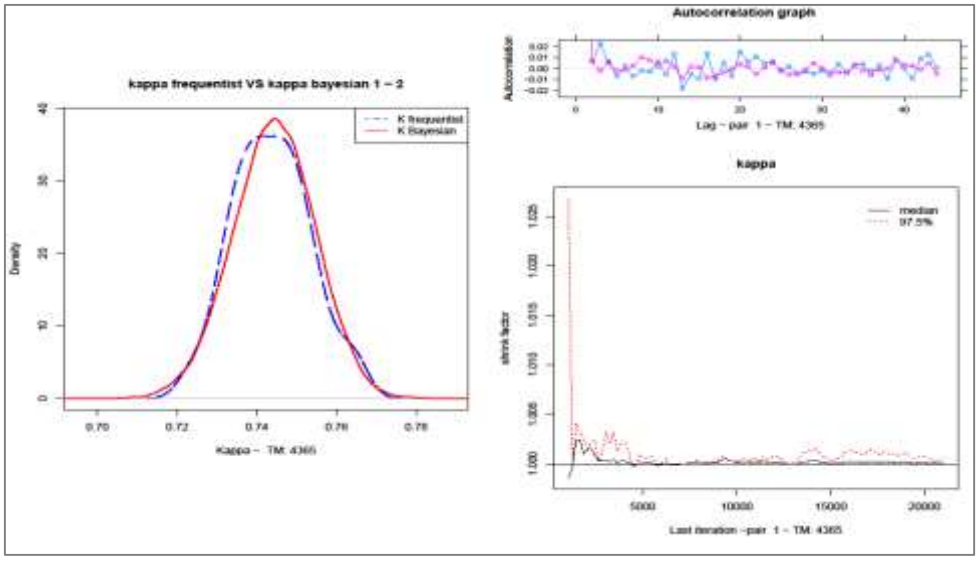
The positive and negative classification of the four cell types analyzed, applying equation (2.19) of the Bayes Theorem, follow the same pattern as that described below for the reactive cells (L_1):

$$\pi(L_1|CT) = \frac{P(L_1)P(CT|L_1)}{P(L_1)P(CT|L_1) + P(L_2)P(CT|L_2) + P(L_3)P(CT|L_3) + P(L_4)P(CT|L_4)}$$

$$\pi(L_1|CF) = \frac{P(L_1)P(CF|L_1)}{P(L_1)P(CF|L_1) + P(L_2)P(CF|L_2) + P(L_3)P(CF|L_3) + P(L_4)P(CF|L_4)}$$

Next, we present the results of the best classifier analyzed with the Dirichlet - Dirichlet model, in this case the Linear Discriminant Analysis (LDA) (see Figure 7). In addition, the final and graphic reports of all the pairs of classifiers are published at: https://github.com/RicardoBorja.

Figure 7. Algorithm results LDA-TRUE



A slight increase in the credibility intervals (K Bayesian) and greater shoring in the posteriori kappa distribution are visualized, considering that a 95% credibility interval was worked on, representing the interval where there is a probability equal to 0.95 that contain kappa. In addition, it is observed that the two chains do not show correlation and converge with a burn-in of 10000.

Two more tests were performed adding information in the a priori distribution, taking into account the prevalence of each leukemic cell at Hospital Clinic level (inside) and Spain (outside). Tables 7 and 8 summarize the results obtained in the estimation of the kappa index only of the gold standard compared to LDA. Final reports and graphs of all pairs of observers can be observed at: https://github.com/RicardoBorja.

Table 7. *Results frequentist method - application to leukemic cells LDA-TRUE.*

FREQUENTIST METHOD					
SAMPLE SIZE	PAIR OF OBSER.	KAPPA LOWER	KAPPA	KAPPA UPPER	ACCURACY
4365	1-2	0.7251	0.7444	0.7639	0.8685

Table 8. *Bayesian method results with prevalence of leukemic cells LDA-TRUE*.

BAYESIAN METHOD							
SAMPLE SIZE	MODEL	PAIR OF OBSER.	KAPPA LOWER	KAPPA	KAPPA UPPER	P-VALUE 2 CHAINS	PREVALENCE
4365	DI-DI	1-2	0.7236	0.7444	0.7641	(0.84;0.94)	NOT
4365	DI-DI	1-2	0.7239	0.7443	0.7644	(0.97;0.14)	INSIDE
4365	DI-DI	1-2	0.7236	0.7422	0.7638	(0.97;0.77)	OUTSIDE
4365	DI-MUL	1-2	0.5451	0.8339	0.8707	(0.31;0.53)	NOT
4365	DI-MUL	1-2	-55.26	0.8264	0.8747	(0.05;0.67)	INSIDE
4365	DI-MUL	1-2	-	-	-	-	OUTSIDE
4365	BE-BE	1-2	0.8582	0.8683	0.8782	(0.51;0.80)	NO
4365	BE-BE	1-2	0.8581	0.8683	0.8783	(0.97;0.80)	INSIDE
4365	BE-BE	1-2	0.8583	0.8683	0.8782	(0.11;0.41)	OUTSIDE

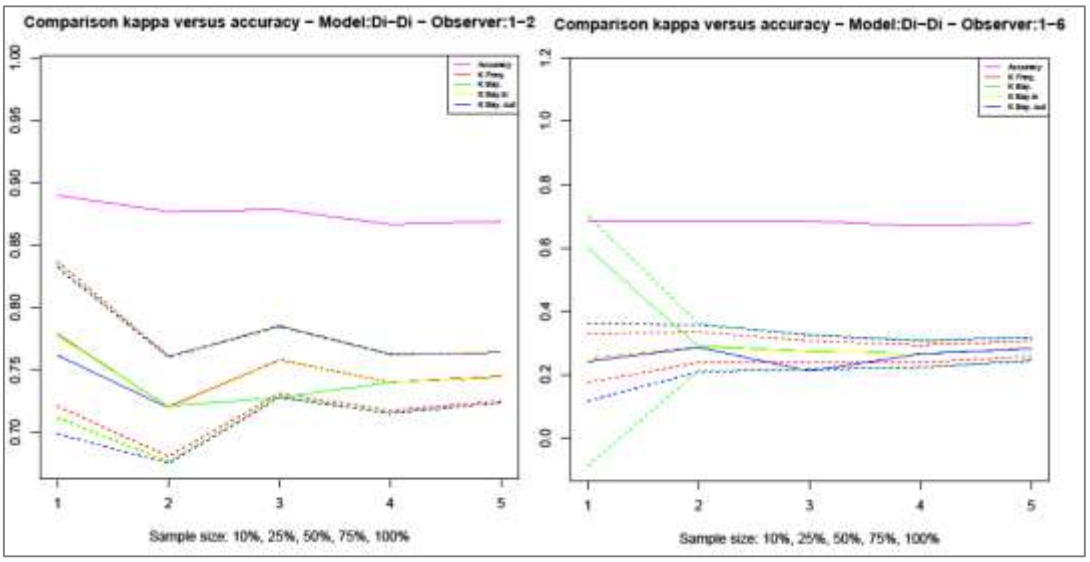
The study showed that the Dirichlet - Dirichlet model was the most optimal and robust in the estimation of the kappa index, demonstrating a high convergence value of its two

chains in all cases of prevalence, especially in the most extreme (outside – p-value = (0.97;0.77)), unlike the other two models. In addition, their credibility intervals become more leptokurtic while the prevalence is more extreme, adjusting kappa effectively for unbalances cases. The percentage of variation of kappa in each model is small due to the high amount of sample data with which we work. The best algorithm was LDA, which presented a good agreement in relation to the gold standard with an observed agreement of 86.8% and expected by chance of 48.55%; while SVM and RF had a moderate agreement.

4.2 Index Kappa versus accuracy

The KFreqBay library was applied using the Dirichlet - Dirichlet model to the algorithm with higher and lower accuracy (LDA and RF), in the leukemic cell database, randomly selecting 10%, 25%, 50%, 75% and 100% of the sample size (4365). We worked under three scenarios: equiprobable, prevalence of the Hospital and of Spain. Figures 8 represent the results obtained.

Figure 8. *Kappa evolution graph and accuracy by sample size – LDA (1-2) y RF (1-6).*



It was confirmed that in both classifiers with high (observer: 1-2) or low precision (observer: 1-6) the kappa index especially Bayesian, when adding a prior information (Bay-in and Bay-out) shows a greater sensitivity to sample change and class proportion. Their credibility intervals increase in relation to the frequentist kappa by adding information, taking into account that in the Bayesian results in both algorithms the more critical the prevalence between them the credibility intervals decrease, thus improving the parameter estimation. However, it is evident that the accuracy is almost invariable,

even more so in algorithms with lower performance. It was shown that the best way to compare classifiers, especially with unbalanced classes, is through Bayesian concordance methods.

5. CONCLUSIONS

The kappa index is a very efficient but underutilized metric, which allows to know the performance of a classifier especially in the case of unbalanced multiclass problems, in comparison with the accuracy, which is a widely used measure but does not provide a complete picture of the performance of the analyzed classifiers.

The Bayesian kappa index (BKI), that we propose, is the optimal tool to evaluate the degree of agreement between two observers or classifiers in the unbalanced multiclass case, due to the correction of the chance effect. It allows enter information of the prevalence within the prior distribution. The three Bayesian models implemented in this paper demonstrated the robustness and sensitivity of the KFreqBay library executable in the R environment with free access. It allows to develop a frequentist and Bayesian concordance analysis either with a pre-established database or through the simulation of a gold standard and observers through a multinomial distribution. When the sample size decreases and the frequencies of the classes are more extreme, the kappa index shows sensitivity and experiences a widening of the credibility intervals. The expected agreement by chance reacts inversely proportional to kappa index.

The Bayesian concordance analysis applied in the case study of the classification of leukemic cells highlights the advantages and effectiveness of the method proposed with the designed library. In fact, the adjustments in the value of kappa under extreme prevalence scenarios allowed us to know the differences when evaluating a classifier depending on the reality to which it is exposed, in this case at the level of the Hospital Clinic or at the level of Spain.

REFERENCES

Blei, D. M., Nigle, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, *3*, 993-1022.

Boldú, L., Merino, A., Alférez, S., Molina, A., Acevedo, A., Rodellar, J. (2019). Automatic recognition of different types of acute leukemia in peripheral blood by images analysis. Journal of clinical Pathology, DOI: 10.1136/jclinpath-2019-205949.

Borja, R. S. (2019). *Método de concordancia bayesiano y su aplicación en problemas de clasificación multiclase con categorías desequilibradas* (Universidad Politécnica de Cataluña - UB). Available at: https://upcommons.upc.edu/handle/2117/127344

Bradley, E. (2013). Bayes' Theorem in the 21st Century. *Science*, *340*(6137), 1177-1178, DOI:10.1126/science.1236536.

Brennan, R. L., & Prediger, D. J. (1981). Coefficient Kappa: Some Uses, Misuses, and Alternatives. *Educational and Psychological Measurement*, *41*(3), 687-699, DOI:10.1177/001316448104100307.

Chib, S., Nardari, F., & Shephard, N. (2002). Markov chain Monte Carlo methods for stochastic volatility models. *Journal of Econometrics*, *108*(2), 281-316, DOI: 10.1016/S0304-4076(01)00137-3.

Cristobal, A. (2000). *Inferencia Estadística* (2da ed.). Zaragoza: Prensas Universitarias de Zaragoza.

D`Agostini, G. (2003). Bayesian inference in processing experimental data: Principles and basic applications. *Reports on Progress in Physics*, *66*(9), 1383–1419, DOI: 10.1088/0034-4885/66/9/201

Fleiss, J., Levin, B., & Cho, M. (2003). *Statistical Methods for Rates and Proportions* (3era ed.). New Jersey: John Wiley & Sons.

Ford, E. B. (2015, junio 5). Convergence Diagnostics For Markov Chain Monte Carlo [online]. Available at: https://astrostatistics.psu.edu/RLectures/diagnosticsMCMC.pdf

Garabedian, C., Butruille, L., Drumez, E., Servan Schreiber, E., Bartolo, S., Bleu, G., … Houfflin-Debarge, V. (2017). Inter-observer reliability of 4 fetal heart rate classifications. *Journal of Gynecology Obstetrics and Human Reproduction*, *46*(2), 131-135, DOI:10.1016/j.jogoh.2016.11.002.

Gupta, A., & Nadarajah, S. (2004). *Handbook of Beta Distribution and Its Applications*. Broken Sound Parkway NW: CRC Press.

Hillier, F., & Lieberman, G. (2010). *Introducción a la Investigación de Operaciones* (9na ed.). México: McGraw Hill.

Koch, K. (1990). *Bayes Theorem* (Vol. 31). Berlin: Springer.

Landis, J. R., & Koch, G. G. (1977). The Measurement of Observer Agreement for Categorical Data. *Biometrics*, *33*(1), 159-174, DOI: 10.2307/2529310.

Laurence, M., & Pascal, M. (2009). Bayesian decision theory as a model of human visual perception: Testing Bayesian transfer. *Visual Neuroscience*, *26*(1), 147-155, DOI:10.1017/S0952523808080905.

Lebreton, J. M., Ployhart, R. E., & Ladd, R. T. (2004). A Monte Carlo Comparison of Relative Importance Methodologies. *Organizational Research Methods*, *7*(3), 258-282, DOI:10.1177/1094428104266017

Maxwell, A. E., Warner, T. A., & Fang, F. (2018). Implementation of machine-learning classification in remote sensing: An applied review: International Journal of Remote Sensing: Vol 39, No 9. *International Journal of Remote Sensing*, *39*(9), 2784-2817, DOI:10.1080/01431161.2018.1433343.

McHugh, M. (2012, 15). Interrater reliability: The kappa statistic. *Biochemia Medica*, *22*(3), 276-282.

Monleón-Getino T, Rodríguez-Casado CI, Verde PE. 2019. Shannon Entropy Ratio, a Bayesian Biodiversity Index Used in the Uncertainty Mixtures of Metagenomic Populations. Journal of Advanced statistics 4(4) 1-23.

Powers, D. M. W. (2012). The Problem with Kappa. *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, 345–355. Stroudsburg, PA, USA: Association for Computational Linguistics.

Press, J. (1989). *Bayesian Statistics principles, models and applications*. Califonia: John Wiley & Sons.

Press, James. (2009). *Subjective and Objective Bayesian Statistics: Principles, Models, and Applications* (2da ed.). New Yersey: John Wiley & Sons.

Sanjib, B., Mousumi, B., & Ananda, S. (2000). Bayesian Inference for Kappa from Single and Multiple Studies. *Biometrics*, *56*(2), 577-582, DOI:10.1111/j.0006-341X.2000.00577.x

Sheldom, R. (2014). *Introduction to Probability Models* (11th ed.). Los Angeles, California: Academic Press.

Shridhar, K., Laumann, F., & Liwicki, M. (2019). A Comprehensive guide to Bayesian Convolutional Neural Network with Variational Inference. *arXiv:1901.02731 [cs, stat]*.

Shuo, W., & Xin, Y. (2012). Multiclass Imbalance Problems: Analysis and Potential Solutions. *IEEE Transactions on Systems*, *42*(4), 1119-1130, DOI:10.1109/TSMCB.2012.2187280

Watson, P., & Petrie, A. (2010). *Method agreement analysis: A review of correct methodology*. *73*, 1167-1179.

Westgard, J. (2008). *Basic method validation* (3ra ed.). Wisconsin: Madison.