An Ensemble of Naive Bayes Classifiers for Uncertain Categorical Data

Marcelo Rodrigues de Holanda Maia*†, Alexandre Plastino* and Alex A. Freitas‡
*Instituto de Computação
Universidade Federal Fluminense, Niterói, RJ, Brazil
E-mail: mmaia@ic.uff.br, plastino@ic.uff.br
†Instituto Brasileiro de Geografia e Estatística, Rio de Janeiro, RJ, Brazil
E-mail: marcelo.h.maia@ibge.gov.br
‡School of Computing
University of Kent, Canterbury, Kent, UK
E-mail: a.a.freitas@kent.ac.uk

***Abstract*—Coping with uncertainty is a very challenging issue in many real-world applications. However, conventional classification models usually assume there is no uncertainty in data at all. In order to fill this gap, there has been a growing number of studies addressing the problem of classification based on uncertain data. Although some methods resort to ignoring uncertainty or artificially removing it from data, it has been shown that predictive performance can be improved by actually incorporating information on uncertainty into classification models. This paper proposes an approach for building an ensemble of classifiers for uncertain categorical data based on biased random subspaces. Using Naive Bayes classifiers as base models, we have applied this approach to classify ageing-related genes based on real data, with uncertain features representing protein-protein interactions. Our experimental results show that models based on the proposed approach achieve better predictive performance than single Naive Bayes classifiers and conventional ensembles.**

***Keywords*-Classification, Ensemble, Uncertain data, Naive Bayes, Bioinformatics**

I. Introduction

Data uncertainty is a common issue in many real-world domains due to various reasons, including measurement errors, data staleness, repeated measurements, data generation and collection process. Coping with uncertainty is a challenging task in data mining applications since the reliability of the information used to build models significantly impacts their performance. However, conventional classification methods usually assume that data are precisely defined, effectively disregarding uncertainty.

In order to fill this gap, there has been a growing number of studies addressing the problem of classification based on uncertain data. Although some methods resort to ignoring uncertainty or artificially removing it from data, strategies for actually incorporating information on uncertainty into classification models have produced promising results, showing that this kind of approach can improve predictive performance [1]–[4].

In this paper, we propose a new approach for building an ensemble of classifiers tailored to cope with uncertainty in the values of categorical features. We rely on the hypothesis that the higher the degree of uncertainty for a given feature, the less it might contribute to the predictive performance as it provides less reliable information about the instances to be classified. We evaluate this proposed approach by applying it to build an ensemble of Naive Bayes classifiers.

Our experiments involve datasets of ageing-related genes containing uncertain features. Ageing can be defined as a progressive decline in the fitness of an organism that occurs with increasing age, ultimately ending in death. While it is unclear precisely what mechanisms drive ageing, genes certainly play an essential role in it [5]. Therefore, ageing genetics is an important subject in computational biology. In addition, ageing is a strategic research area because the proportion of elderly individuals among the population is increasing fast, and old age is the greatest risk factor for many diseases (including, e.g., most types of cancer).

The remainder of this paper is organized as follows. Section II reviews the background on the main topics covered in this work. In Section III, we introduce our proposed novel approach. Section IV describes our experimental methodology. Section V reports the results obtained. Finally, we present conclusions in Section VI.

II. Background

A. Ensemble Methods

In this work, we consider ensemble methods categorized as averaging methods. They build several base classifiers independently on random subsets of the original training set. Then, they aggregate the individual base classifiers' predictions to form a combined prediction.

These methods can be differentiated by how they draw random subsets of the original training set. In particular, Bagging methods are those that draw random subsets of the instances in the dataset with replacement [6], and if a method

This study was financed in part by: Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq, Brazil) [grant number 310444/2018-7]; Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES, Brazil) [finance code 001]; and Instituto Brasileiro de Geografia e Estatística (IBGE, Brazil).

draws random subsets of the features in the dataset, then it is known as Random Subspaces [7].

B. Naive Bayes Classifiers

Given a class variable y and a feature vector $X = (x_1, x_2, \ldots, x_m)$, based on the application of Bayes' theorem under the "naive" assumption that the features are conditionally independent given the value of the class variable, a Naive Bayes classifier predicts the class y that maximizes the approximation of $P(y|X)$ given by:

$$P(y|X) \propto P(y) \prod_{j=1}^{m} P(x_j|y) \quad (1)$$

We have chosen Naive Bayes as the base classifier to evaluate our proposed ensemble scheme since it has obtained good results in many real-world domains, including the classification of ageing-related genes [8]–[10] – the target application domain in this work. Furthermore, there are several reports on the use of Naive Bayes classifiers to build ensembles in the literature [11]–[13].

C. Classification with Uncertain Data

Data uncertainty is a common issue in many real-world applications due to various reasons, including measurement errors, data staleness, data generation and collection process. The forms of data uncertainty have been usually classified into existential uncertainty or value uncertainty.

Existential uncertainty refers to the case when it is uncertain whether an object exists. For example, an instance in a dataset could be associated with a probability representing the confidence of its occurrence.

In contrast, value uncertainty, which we address in this work, refers to the case when an instance is known to exist, but its feature values are not precisely known. An uncertain feature value is usually represented by a probability distribution on the domain of the feature.

Uncertainty in numerical feature values has been the focus of several studies, using multiple types of classification models such as neural networks [1], decision trees [2], k-nearest neighbors [3] and support vector machines [4]. Although some of those approaches could be adapted to cope with uncertain categorical features, this kind of value uncertainty has received relatively little attention in the literature. Another noticeable characteristic about past work in this field is that most experiments have been performed on data that were not originally uncertain. Instead, uncertainty was artificially introduced into the data through augmentation processes. In contrast, we evaluate our proposed approach using real-world data with uncertain categorical features.

III. Proposed Approach

The specification of the approach proposed in this work considers the following definitions.

Let $F = \{f_1, f_2, \ldots, f_m\}$ be the set of predictive features, where $m \geq 1$, and $C = \{c_1, c_2, \ldots, c_q\}$ be the set of classes, where $q \geq 2$. The domain of a feature f_j is $dom(f_j)$. A dataset $D = \{(X_1, y_1), (X_2, y_2), \ldots, (X_n, y_n)\}$ consists of n labelled instances. Each instance in D, identified by an index i, is associated with a feature vector $X_i = (x_{i1}, x_{i2}, \ldots, x_{im})$ and a class label $y_i \in C$. The classification problem is to construct a model from D that is capable of predicting the class of an unlabelled instance given its corresponding feature vector.

Our uncertainty framework considers that some of the features are uncertain, i.e., there is a set of uncertain features $U \subseteq F$, all of which are assumed to be categorical. If f_j is a categorical feature, its domain is a finite set of values $dom(f_j) = \{v_{j1}, v_{j2}, \ldots, v_{j|dom(f_j)|}\}$, $|dom(f_j)| \geq 2$. If a feature f_j is not uncertain, its corresponding value x_{ij} for an instance i is represented by a single value v_{ij}. Otherwise it is a discrete probability distribution represented by a probability vector P_{ij}. That is:

$$x_{ij} = \begin{cases} v_{ij} \in dom(f_j), & \text{if } f_j \in F \setminus U \\ P_{ij} = (p_{ij1}, p_{ij2}, \ldots, p_{ij|dom(f_j)|}), & \text{otherwise} \end{cases}$$

where, if $f_j \in U$, $p_{ijk} \in [0, 1]$ represents the probability that x_{ij} assumes the value v_{jk} and $\sum_{k=1}^{|dom(f_j)|} p_{ijk} = 1$.

A. An Ensemble Approach for Coping with Uncertainty in Categorical Features

We propose a new approach for building an ensemble of classifiers that incorporates uncertainty about the value of categorical features into the model. The intuition motivating this proposal is that the higher the degree of uncertainty for a given feature, the less it might contribute to the predictive performance as it provides less reliable information about the instances to be classified. Furthermore, missing values, i.e., the absence of values for a feature in a dataset, are also considered since they represent another factor that may undermine the contribution of a feature to the model.

This approach relies on the use of a bias value computed for each feature f_j based on its degree of uncertainty and on its fraction of missing values in the dataset, given by:

$$b_j = \left(1 - \frac{1}{|I \setminus M_j|} \sum_{i \in I \setminus M_j} E_{ij}\right) \times \frac{|I \setminus M_j|}{|I|}$$

where $I = \{1, 2, \ldots, n\}$ is the set of indices of all instances in D, M_j is the set of indices of instances in D with a missing value for the feature f_j, and E_{ij} is the entropy of the probability distribution represented by P_{ij} if f_{ij} is an uncertain feature (or zero, otherwise), that is:

$$E_{ij} = \begin{cases} -\sum_{k=1}^{|dom(f_j)|} p_{ijk} log(p_{ijk}), & \text{if } f_j \in U \\ 0, & \text{otherwise} \end{cases}$$

In the feature bias definition, the first factor (between parentheses) is the complement of the mean entropy over

all probability distributions associated with the feature f_j, whereas the second factor is the fraction of non-missing values for the feature. Therefore, the computed bias is a value in the range $[0, 1]$ with higher values indicating lower uncertainty degrees, i.e., more reliable features.

The feature bias values are normalized over all features, defining a probability distribution $B = (\beta_1, \beta_2, \ldots, \beta_m)$, where a probability β_j associated with a feature f_j is given by $\beta_j = b_j / (\sum_{l=1}^{m} b_l)$.

Recall that in the general Random Subspaces strategy, each base classifier in the ensemble is trained with a different set of features, sampled from the full set F. In this approach, we use the probability distribution B to sample the features to be considered by each base classifier in the ensemble instead of the default uniform distribution. Hence, we call our proposed approach Biased Random Subspaces (BRS).

Note that no assumption is made about if or how the base classifiers in the ensemble handle uncertain data, as our focus is on the BRS approach to cope with uncertainty at the ensemble level. Even if the base classifiers do not cope with uncertainty, this approach can still be straightforwardly applied. As an example, it can be done by replacing each probability distribution P_{ij} corresponding to an uncertain feature in the dataset with its expected value, i.e., the value v_{jk} that maximizes p_{ijk}.

IV. Experimental Methodology

A. Dataset Creation

In this work, we have applied the proposed approach to the classification of ageing-related genes in different types of organisms, which several studies have addressed in recent years [8]–[10], [14]–[16]. In this problem, the objective is to identify the effect of genes on the longevity of an organism. More specifically, given an ageing-related gene, the problem is to predict whether its effect on the lifespan of an organism is positive (pro-longevity) or negative (anti-longevity).

The GenAge database, part of the Human Ageing Genomic Resources (HAGR) collection [17], is an essential resource in this context, comprising data about over two thousand genes, including their classification regarding longevity influence.

Genes encode proteins, and information about the interaction between two proteins, known as protein-protein interaction (PPI), has been used in past work addressing the classification of ageing-related genes [10], [14], [16], leading to improvements in predictive performance. STRING [18] is a database of PPIs that stem from computational predictions, from knowledge transfer between organisms, and from interactions aggregated from other databases.

Due to the technical difficulties of detecting PPIs via biological experiments, the available information on PPI is incomplete and exhibits varying levels of reliability. Furthermore, it is complemented with computational predictions (which are less reliable than biological experiments in general). Therefore, the STRING database provides a score for each PPI, computed by combining the probabilities from the different evidence channels, which means the available data on PPI are intrinsically uncertain.

We have generated four datasets[1] of ageing-related genes by integrating data from the GenAge database (Build 20) and the STRING database (Version 11.0). Each dataset contains data regarding ageing-related genes of one of the four major biomedical model organisms from the GenAge database: *C. elegans* (roundworm), *D. melanogaster* (fruit fly), *M. musculus* (mouse), and *S. cerevisiae* (baker's yeast).

Each instance in our datasets refers to an ageing-related gene of the corresponding model organism and consists of uncertain features referring to PPIs and a binary class variable indicating if the instance is positive (pro-longevity gene) or negative (anti-longevity gene) according to the GenAge database. Each PPI feature refers to one protein and has a binary domain, indicating whether or not an interaction between the protein encoded by the corresponding gene (the current instance) and the protein referred by the feature has been observed. Since these features are uncertain, they are represented by probability distributions according to our uncertainty framework.

A value x_{ij} for a PPI feature f_j corresponding to an instance i in the dataset is represented by a probability distribution $P_{ij} = (p_{ij1}, p_{ij2})$, where p_{ij1} and p_{ij2} are the complimentary probabilities of x_{ij} assuming each of the two values in $dom(f_j)$. Therefore, each probability distribution associated with a PPI feature value can actually be encoded by a single value p_{ij}, in which case $P_{ij} = (p_{ij}, 1 - p_{ij})$. In our datasets, this value is the confidence score obtained from the STRING database for the corresponding PPI, which indicates its probability of occurrence.

Some distinctive characteristics of these datasets, which make them quite challenging, are a very large number of features, a small number of instances, and a very high percentage of missing values (which occur when there is no information regarding a specific PPI in the STRING database). We have discarded PPI features with low support (annotating less than ten genes) to avoid overfitting.

Table I presents detailed information about the datasets. The first column indicates the corresponding model organism, whereas the remaining columns present, respectively, the number of instances, the number of features, the percentage of missing values, and the percentage of instances corresponding to each class (Anti- and Pro-longevity).

B. Algorithms Being Evaluated

In the experiments, we consider two baseline methods, both based on conventional Naive Bayes classifiers. Each of these methods uses a different interpretation of the uncertain feature values since they do not cope with uncertainty.

[1]The datasets used in the experiments are publicly available on the web at https://github.com/marcelorhmaia/ensembles-for-uncertain-data

Table I
INFORMATION ABOUT THE DATASETS USED IN THE EXPERIMENTS

Dataset	Instances	Features	Missing Values (%)	Class (%)	
				Anti	Pro
C. elegans	763	9692	93.8	66.3	33.7
D. melanogaster	185	3883	88.4	37.3	62.7
M. musculus	82	4216	78.4	37.8	62.2
S. cerevisiae	382	4274	90.3	88.0	12.0

The first baseline method (referred to as NB-NV) treats each uncertain value (the probability of occurrence of the corresponding PPI) as a numeric value. Therefore, it assumes that $dom(f_j) = [0, 1], \forall j \in \{1, 2, \ldots, m\}$ and that the feature value probability distributions are Gaussian.

The other baseline method (referred to as NB-EV) replaces each uncertain value with the expected value from the corresponding probability distribution, i.e., it binarizes the value representing the probability of occurrence of the corresponding PPI using the threshold value 0.5. It considers multivariate Bernoulli distributions for the data.

In both baseline methods, we replace missing values with zeros, i.e., if there is no information regarding a PPI in the dataset, we assume it does not occur, a case represented by a zero value, as usual in the literature using PPIs as predictive features for classifying genes.

For each baseline method, we build an ensemble that uses the conventional Bagging and Random Subspaces strategies (referred to as ENB-NV and ENB-EV, respectively) and another one that uses our proposed BRS approach in combination with conventional Bagging (ENB-NV+BRS and ENB-EV+BRS, respectively). Note that ENB-NV and ENB-EV do not cope with uncertainty, but only the proposed ENB-NV+BRS and ENB-EV+BRS ensembles do so.

We have used available implementations from the scikit-learn library [19] as the baseline methods, as well as for the conventional ensembles. The algorithms based on our proposed approaches have been implemented through the extension of scikit-learn's original methods[2]. We have set the ensembles to use 500 base classifiers, building each one on a different subset of the training set consisting of n instances drawn by the Bagging procedure and $\sqrt{m}$ features drawn by the Random Subspaces (or the BRS) procedure.

We first separate the experiments into two groups (consisting of methods based on the NV and EV approaches), and then compare three algorithms in each group:

- a single Naive Bayes classifier (one of the two baseline methods)
- a conventional ensemble of this base classifier
- another ensemble that uses our proposed BRS approach

Then we compare the best models from the two groups.

[2]The source-code used in the experiments is publicly available on the web at https://github.com/marcelorhmaia/ensembles-for-uncertain-data

C. Measuring Predictive Performance

We assess the predictive performance of the evaluated algorithms using two metrics: the Area Under the Receiver Operating Characteristic curve (AUROC) and the geometric mean of sensitivity and specificity.

The ROC curve is a method for evaluating the performance of a binary classifier by plotting its true-positive rate (sensitivity) versus its false-positive rate (one minus the specificity) at various threshold settings. The AUROC summarizes this information into one number.

The geometric mean of sensitivity and specificity (G-mean) measures the balance between predictive performances on both the majority and minority classes. Therefore, it is suitable for assessing predictive performance on imbalanced datasets, like the ones used in our experiments.

Each algorithm was evaluated by running a well-known 10-fold cross-validation procedure. In addition, we have assessed the statistical significance of the differences in the predictive performance measures between each pair of algorithms. We have used a paired Wilcoxon signed-rank test for each dataset for this evaluation, considering a significance level of 0.05.

V. COMPUTATIONAL RESULTS

In the first experiment, we have compared the three models based on NB-NV. Table II presents the specificity (Spec.) and sensitivity (Sens.) values obtained by each model, as these measures are used to compute predictive performance metrics. Specificity and sensitivity correspond to the recall values for the Anti-longevity and Pro-longevity classes, respectively.

Table II
SPECIFICITY AND SENSITIVITY VALUES (%) FOR MODELS BASED ON NB-NV

Dataset	NB-NV		ENB-NV		ENB-NV+BRS	
	Spec.	Sens.	Spec.	Sens.	Spec.	Sens.
C. elegans	70.54	56.61	35.54	94.47	42.30	87.47
D. melanogaster	36.19	85.12	76.65	42.85	80.23	44.56
M. musculus	46.83	86.45	64.33	50.67	64.33	49.24
S. cerevisiae	99.69	0.00	50.68	65.83	40.81	85.00

Tables III and IV present the results for this group of models regarding the AUROC and G-mean metrics, respectively. Besides the predictive performance values achieved by each model on each dataset, these tables present the average rank for each model in the last row. The best value in each comparison is presented in bold. Furthermore, the statistically significant differences in the comparisons between the model using our proposed BRS approach and each of the other models are indicated by superscript symbols next to the respective higher values.

Based on the results presented in Tables III and IV, the general conclusion from this first experiment is that ENB-

Table III
AUROC RESULTS (%) FOR MODELS BASED ON NB-NV

Dataset	NB-NV	ENB-NV	ENB-NV+BRS
C. elegans	63.52	71.46	**72.33**[a]
D. melanogaster	60.83	**65.62**	65.03
M. musculus	67.93	66.73	**68.51**
S. cerevisiae	49.84	**61.62**	61.22[a]
Avg. Rank	2.75	1.75	**1.50**

[a]Statistically significant (NB-NV vs. ENB-NV+BRS), p-value = 0.003 for *C. elegans*, p-value = 0.004 for *S. cerevisiae*.

Table IV
G-MEAN RESULTS (%) FOR MODELS BASED ON NB-NV

Dataset	NB-NV	ENB-NV	ENB-NV+BRS
C. elegans	**63.19**	57.94	60.83[b]
D. melanogaster	55.50	57.31	**59.79**
M. musculus	**63.63**	57.09	56.28
S. cerevisiae	0.00	57.76	**58.90**[a]
Avg. Rank	2.00	2.25	**1.75**

[a]Statistically significant (NB-NV vs. ENB-NV+BRS), p-value = 0.003.
[b]Statistically significant (ENB-NV vs. ENB-NV+BRS), p-value = 0.033.

NV+BRS is the best model regarding both AUROC and G-mean, with the average ranks of 1.50 and 1.75, respectively.

In the second experiment, we have compared the three models based on NB-EV. Table V presents the specificity and sensitivity values obtained by each model, whereas Tables VI and VII present the results for this group of models regarding the AUROC and G-mean metrics, respectively.

Table V
SPECIFICITY AND SENSITIVITY VALUES (%) FOR MODELS BASED ON NB-EV

Dataset	NB-EV		ENB-EV		ENB-EV+BRS	
	Spec.	Sens.	Spec.	Sens.	Spec.	Sens.
C. elegans	74.99	58.42	96.44	13.64	90.27	26.89
D. melanogaster	26.91	86.26	8.33	96.09	8.33	94.55
M. musculus	34.83	85.14	18.67	92.64	23.67	92.64
S. cerevisiae	85.56	36.67	98.83	5.00	94.45	24.17

Table VI
AUROC RESULTS (%) FOR MODELS BASED ON NB-EV

Dataset	NB-EV	ENB-EV	ENB-EV+BRS
C. elegans	76.89[a]	**76.91**[b]	74.81
D. melanogaster	66.65	64.05	**69.00**
M. musculus	66.87	69.26	**69.30**
S. cerevisiae	**77.13**	75.55	76.79
Avg. Rank	2.00	2.25	**1.75**

[a]Statistically significant (NB-EV vs. ENB-EV+BRS), p-value = 0.012.
[b]Statistically significant (ENB-EV vs. ENB-EV+BRS), p-value = 0.033.

Based on the results presented in Tables VI and VII, the general conclusion from the second experiment is that

Table VII
G-MEAN RESULTS (%) FOR MODELS BASED ON NB-EV

Dataset	NB-EV	ENB-EV	ENB-EV+BRS
C. elegans	**66.19**[a]	36.27	49.27[b]
D. melanogaster	**48.18**[a]	28.28	28.06
M. musculus	**54.46**	41.59	46.82
S. cerevisiae	**56.01**	22.23	47.78[b]
Avg. Rank	**1.00**	2.75	2.25

[a]Statistically significant (NB-EV vs. ENB-EV+BRS), p-value = 0.003 for *C. elegans*, p-value = 0.010 for *D. melanogaster*.
[b]Statistically significant (ENB-EV vs. ENB-EV+BRS), p-value = 0.005 for *C. elegans*, p-value = 0.017 for *S. cerevisiae*.

the proposed ENB-EV+BRS is the best model regarding AUROC and the second best (out of three) regarding G-mean, with the average ranks of 1.75 and 2.25, respectively. In this experiment, a single NB-EV classifier performed better than both ensembles regarding G-mean. Nonetheless, it is noticeable that our proposed BRS approach was still able to improve the predictive performance of an ensemble, as the ENB-EV+BRS model outperformed the ENB-EV.

Our third experiment aims at determining the best overall method regarding the AUROC metric. Table VIII presents the results for ENB-NV+BRS and ENB-EV+BRS, the best models from experiments 1 and 2, respectively, regarding this metric. The proposed ENB-EV+BRS was the best overall model in this comparison, outperforming the ENB-NV+BRS for all datasets.

Table VIII
AUROC RESULTS (%) FOR THE BEST MODEL FROM TABLE III AND THE BEST MODEL FROM TABLE VI

Dataset	ENB-NV+BRS[a]	ENB-EV+BRS[b]
C. elegans	72.33	**74.81**
D. melanogaster	65.03	**69.00**
M. musculus	68.51	**69.30**
S. cerevisiae	61.22	**76.79***
Avg. Rank	2.00	**1.00**

[a]Results from Table III. [b]Results from Table VI.
*Statistically significant, p-value = 0.012.

Finally, our fourth experiment aims at determining the best overall method regarding the G-mean metric. Table IX presents the results for ENB-NV+BRS and NB-EV, the best models from experiments 1 and 2, respectively, regarding this metric. ENB-NV+BRS was the best overall model in this comparison, with an average rank of 1.25.

As a general conclusion from the reported experiments, we can point out that the results support the hypothesis that our proposed BRS approach improves the predictive performance of ensembles on uncertain data, as the best overall models for the AUROC and G-mean metrics were the ENB-NV+BRS and ENB-EV+BRS, respectively, both based on the BRS approach.

Table IX
G-MEAN RESULTS (%) FOR THE BEST MODEL FROM TABLE IV AND THE BEST MODEL FROM TABLE VII

Dataset	ENB-NV+BRS[a]	NB-EV[b]
C. elegans	60.83	**66.19**
D. melanogaster	**59.79***	48.18
M. musculus	**56.28**	54.46
S. cerevisiae	**58.90**	56.01
Avg. Rank	**1.25**	1.75

[a]Results from Table IV. [b]Results from Table VII.
*Statistically significant, p-value = 0.038.

VI. CONCLUSIONS

In this work, we have addressed the problem of classification in datasets containing categorical features with uncertain values, i.e., values represented by probability distributions in the respective features' domains. We have proposed an ensemble approach called Biased Random Subspaces (BRS) for coping with this kind of uncertainty, based on the hypothesis that features with lower uncertainty degrees have better class-discrimination power since there is higher confidence about their actual values across the dataset.

Our experiments have compared two types of single Naive Bayes classifiers, conventional ensembles of these classifiers and ensembles based on the BRS approach. We have applied them to classify ageing-related genes from four model organisms based on real data containing uncertain features referring to protein-protein interactions. The results show that the ensembles applying our BRS approach achieved the best overall predictive performance, supporting the hypothesis that applying BRS-based ensembles of classifiers is an effective approach to cope with uncertainty in categorical features, leading to higher predictive performance.

Some directions for future work include applying the proposed BRS approach to build ensembles of other base classifiers than Naive Bayes and perform experiments with other datasets containing uncertain data.

REFERENCES

[1] J. Ge, Y. Xia, and C. Nadungodage, "UNN: A neural network for uncertain data classification," in *Advances in Knowledge Discovery and Data Mining*, 2010, pp. 449–460.

[2] S. Tsang, B. Kao, K. Y. Yip, W.-S. Ho, and S. D. Lee, "Decision trees for uncertain data," *IEEE Trans. Knowl. Data Eng.*, vol. 23, no. 1, pp. 64–78, 2011.

[3] F. Angiulli and F. Fassetti, "Nearest neighbor-based classification of uncertain data," *ACM Trans. Knowl. Discov. Data*, vol. 7, no. 1, 2013.

[4] Z. Xie, Y. Xu, and Q. Hu, "Uncertain data classification with additive kernel support vector machine," *Data Knowl. Eng.*, vol. 117, pp. 87–97, 2018.

[5] D. Wieser, I. Papatheodorou, M. Ziehm, and J. M. Thornton, "Computational biology for ageing," *Philos. Trans. Royal Soc. B: Biol. Sci.*, vol. 366, no. 1561, pp. 51–63, 2011.

[6] L. Breiman, "Bagging predictors," *Mach. Learn.*, vol. 24, no. 2, pp. 123–140, 1996.

[7] T. K. Ho, "The random subspace method for constructing decision forests," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, no. 8, pp. 832–844, 1998.

[8] C. Wan and A. Freitas, "Prediction of the pro-longevity or anti-longevity effect of caenorhabditis elegans genes based on bayesian classification methods," in *IEEE Int. Conf. on Bioinformatics and Biomedicine*, 2013, pp. 373–380.

[9] P. N. da Silva, A. Plastino, and A. A. Freitas, "A novel genetic algorithm for feature selection in hierarchical feature spaces," in *SIAM Int. Conf. on Data Mining*, 2018, pp. 738–746.

[10] P. N. da Silva, A. Plastino, F. Fabris, and A. A. Freitas, "A novel feature selection method for uncertain features: An application to the prediction of pro-/anti- longevity genes," *IEEE/ACM Trans. Comput. Biol. Bioinform.*, 2020.

[11] E. Bauer and R. Kohavi, "An empirical comparison of voting classification algorithms: Bagging, boosting, and variants," *Mach. Learn.*, vol. 36, no. 1, pp. 105–139, 1999.

[12] A. Tsymbal, S. Puuronen, and D. W. Patterson, "Ensemble feature selection with the simple bayesian classification," *Inf. Fusion*, vol. 4, no. 2, pp. 87–100, 2003.

[13] A. Prinzie and D. Van den Poel, "Random multiclass classification: Generalizing random forests to random MNL and random NB," in *Database and Expert Systems Applications*, 2007, pp. 349–358.

[14] T. Huang et al., "Deciphering the effects of gene deletion on yeast longevity using network and machine learning approaches," *Biochim.*, vol. 94, no. 4, pp. 1017–1025, 2012.

[15] C. Wan, A. A. Freitas, and J. P. de Magalhães, "Predicting the pro-longevity or anti-longevity effect of model organism genes with new hierarchical feature selection methods," *IEEE/ACM Trans. Comput. Biol. Bioinform.*, vol. 12, no. 2, pp. 262–275, 2015.

[16] I. Martire, P. N. da Silva, A. Plastino, F. Fabris, and A. A. Freitas, "A novel probabilistic Jaccard distance measure for classification of sparse and uncertain data," in *5th Symp. on Knowledge Discovery, Mining and Learning*, 2017, pp. 81–88.

[17] R. Tacutu et al., "Human Ageing Genomic Resources: new and updated databases," *Nucleic Acids Res.*, vol. 46, no. D1, pp. D1083–D1090, 2017.

[18] D. Szklarczyk et al., "STRING v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets," *Nucleic Acids Res.*, vol. 47, no. D1, pp. D607–D613, 2018.

[19] F. Pedregosa et al., "Scikit-learn: Machine learning in Python," *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, 2011.