Contextualizing Naive Bayes Predictions

Marcelo Loor[1,2(✉)] and Guy De Tré[1]

[1] Department of Telecommunications and Information Processing, Ghent University, Sint-Pietersnieuwstraat 41, 9000 Ghent, Belgium
{Marcelo.Loor,Guy.DeTre}@UGent.be
[2] Department of Electrical and Computer Engineering, ESPOL University, Campus Gustavo Galindo V., Km. 30.5 Via Perimetral, Guayaquil, Ecuador

Abstract. A classification process can be seen as a set of actions by which several objects are evaluated in order to predict the class(es) those objects belong to. In situations where transparency is a necessary condition, predictions resulting from a classification process are needed to be interpretable. In this paper, we propose a novel variant of a naive Bayes (NB) classification process that yields such interpretable predictions. In the proposed variant, augmented appraisal degrees (AADs) are used for the contextualization of the evaluations carried out to make the predictions. Since an AAD has been conceived as a mathematical representation of the connotative meaning in an experience-based evaluation, the incorporation of AADs into a NB classification process helps to put the resulting predictions in context. An illustrative example, in which the proposed version of NB classification is used for the categorization of newswire articles, shows how such contextualized predictions can favor their interpretability.

Keywords: Explainable artificial intelligence · Augmented appraisal degrees · Naive Bayes classification · Context handling

1 Introduction

Computer applications like scoring tools that make judgments about individuals, or graphical applications that incorporate scene recognition to get stunning photos, can be driven by *artificial intelligence* (AI). Although such systems can be very convenient, they might be restricted or avoided in situations where transparency and accountability are highly important. For example, systems that predict the degree to which individuals are suitable (or unsuitable) for a job without explaining their predictions can be banned from using in the European Union according to the *General Data Protection Regulation* (GDPR) [6]. An ongoing challenge in this regard is to find appropriate mechanisms to explain such predictions.

In a previous work [14], we proposed a method to address that challenge in predictions made by a *support vector machine* (SVM) classification process [20,21]. In that method, an evaluation performed to predict whether an object

M.-J. Lesot et al. (Eds.): IPMU 2020, CCIS 1239, pp. 814–827, 2020.
https://doi.org/10.1007/978-3-030-50153-2_60

belongs to a given class or not is augmented in such a way that the object's features supporting the evaluation are also recorded. It has been shown how such an augmentation, which is represented by means of an *augmented appraisal degree* (AAD) [12], can favor the interpretability of SVM predictions.

As a sequel to [14], in this paper we propose a novel version of a naive Bayes (NB) classification process [9], in which AADs are incorporated to contextualize the evaluations performed to predict the class(es) an object belongs to. Our motivation here is that, while the context of evaluations performed by a person can sometimes be inferred from factors like situational or environmental aspects, the context of evaluations carried out by a machine might be difficult to infer. Thus, an explicit representation of the context of evaluations through AADs can help a *NB classifier* (NBC) to offer predictions that are better interpretable.

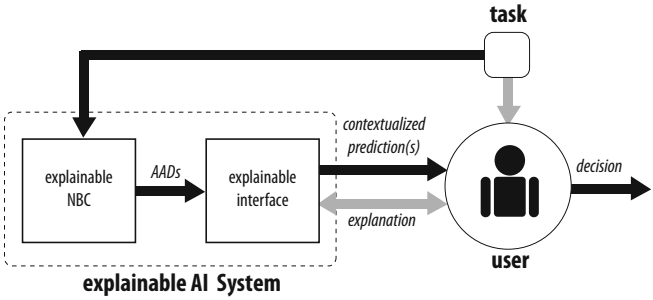

Fig. 1. A general view of the proposed version of NBC in the context of the explanation framework included in the 2016 DARPA report [5].

Contextualized predictions can be useful in situations where informed decisions are needed. In this regard, the proposed NBC, named *explainable NBC* (XNBC), can be included within an *explainable artificial intelligence* (XAI) system [5], by which a user can receive those contextualized predictions to make a decision as shown in Fig. 1. In addition, contextualized predictions can provide direct insights about what is deemed to be relevant to the (knowledge) model used by a classification process. This means that such contextualized predictions can also be used by, say, an AI practitioner to assess the quality of models that result from different learning scenarios.

To illustrate how the novel XNBC works, we develop a *text categorization* process (cf. [10]) by which newswire articles included in the *Reuters-21578* collection [8] are evaluated to predict the class(es) those articles belong to. Figure 2 shows a resulting visual representation where it is indicated why and why not XNBC predicts that a newswire article belongs to a given class up to a specific level: while the size of a word denotes its influence on the classification, its typographical style denotes whether the word is *in favor of* or against the membership in that class. The evaluation behind such a visual representation can also be used by, say, the explainable interface of an XAI system to provide

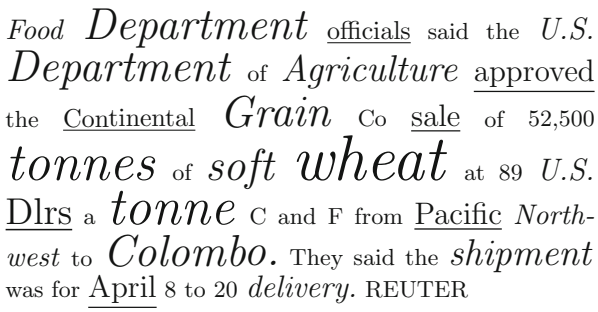

Fig. 2. A visual representation of the reasons that justify why and why not a newswire article belongs to the category 'wheat' up to a specific level.

the following explanation: *"While words like 'Dlrs' or 'Pacific' suggest that the newswire article does not belong to the category 'wheat' with a computed overall grade of* 0.29*, words like 'wheat' or 'tonnes' suggest that the article belongs to the category with a computed overall grade of* 0.71*. These results indicate that the article should be considered member of the category up to a* 0.42*-level."* Notice how this explanation clarifies what has been relevant to the knowledge model used for this prediction.

In the next section, we outline how an integration of the AAD concept into the *intuitionistic fuzzy set* [2,3] concept can be used for the characterization of the evaluation represented in Fig. 2. Then, we describe our novel variant of NBC in Sect. 3 and illustrate how it works in Sect. 4. After that, we present some related work in Sect. 5. The paper is concluded in Sect. 6.

2 Preliminaries

As previously stated, a classification can be seen as a process in which one or more objects are evaluated in order to predict whether those objects can be situated in one or more classes. In situations where an object, say x, has features suggesting that it partially belongs to a given class, say A, a classification algorithm can use the framework of *fuzzy set theory* [23] to model in mathematical terms the evaluation of the level to which x is a member of A. In this framework, such an evaluation can be characterized by a *membership grade*, which is a number $\mu_A(x)$ in the unit interval $[0, 1]$, where 0 and 1 represent in that order the lowest and the highest membership grades. For example, if the newswire article shown in Fig. 2 is denoted by x, and (what has been learned about) the category 'wheat' is represented by A, then $\mu_A(x)$ indicates the level to which x belongs to A. In this regard, if another category, say 'corn', is denoted by B, the expression $\mu_B(x) < \mu_A(x)$ indicates that the level to which (the newswire article) x belongs to B is less than the level to which x belongs to A.

An object can also have features suggesting that it does not belong to a class. Notice in Fig. 2 that, while words such as 'wheat' or 'grain' are in favor of the

membership in the category 'wheat', words like 'Dlrs' or 'Pacific' are against that membership. In this case, a classification algorithm can make use of the *intuitionistic fuzzy set* (IFS) [2,3] framework to model the evaluation of an object x by means of an *IFS element*. An IFS element, say $\langle x, \mu_A(x), \nu_A(x)\rangle$, is constituted by the evaluated object x, a *membership grade* $\mu_A(x)$ and a *nonmembership grade* $\nu_A(x)$, where $\mu_A(x)$ and $\nu_A(x)$ are two numbers in the unit interval $[0, 1]$ that satisfy the *consistency condition* $0 \leq \mu_A(x) + \nu_A(x) \leq 1$. The *buoyancy* [15] of $\langle x, \mu_A(x), \nu_A(x)\rangle$, i.e., $\rho_A(x) = \mu_A(x) - \nu_A(x)$, can be used for comparing this element to another. For example, if $\langle x, \mu_A(x), \nu_A(x)\rangle$ and $\langle x, \mu_B(x), \nu_B(x)\rangle$ denote the evaluations of the membership and nonmembership of x in categories A and B respectively, the expression $\rho_A(x) > \rho_B(x)$ will suggest that x belongs to a larger extent to A than to B.

As can be noticed, neither a membership grade, nor an IFS element can be used to record the object's characteristics that lead to the level to which the object belongs or not to a given class. To record those characteristics, the notion of *augmented appraisal degrees* (AADs) has been proposed in [12]. An AAD, say $\hat{\mu}_{A@K}(x)$, is a pair $\langle \mu_{A@K}(x), F_{\mu_{A@K}}(x)\rangle$ that represents the level $\mu_{A@K}(x)$ to which x belongs to A, as well as the particular collection of x's features $F_{\mu_{A@K}}(x)$ that have been taken into account to determine (the value of) $\mu_{A@K}(x)$ based on the knowledge K. Here, $A@K$ denotes what has been learned about A after following a learning process that yields K as a result. For example, consider that A and x denote the category 'wheat' and the newswire article shown in Fig. 2 respectively. With this consideration, one can use an AAD, say $\hat{\mu}_{A@K}(x) = \langle \mu_{A@K}(x), F_{\mu_{A@K}}(x)\rangle$, to represent the evaluation of the proposition 'x is member of A' according to what has been learned about the category 'wheat' after following a learning process that produces K as a result. In this case, while $\mu_{A@K}(x)$ represents the level to which x belongs to the category 'wheat', $F_{\mu_{A@K}}(x)$ represents the collection of x's words such as 'agriculture', 'grain', or 'wheat' that have been considered for quantifying the value of $\mu_{A@K}(x)$ according to (the knowledge) K.

As has been mentioned above, the newswire article x can also contain words suggesting that it does not belong to the category 'wheat'. To characterize the context of this kind of evaluations, the idea of an *augmented IFS element*, say $\langle x, \hat{\mu}_{A@K}(x), \hat{\nu}_{A@K}(x)\rangle$, has been introduced in [12]. As noticed, an augmented IFS element consists of a membership AAD, $\hat{\mu}_{A@K}(x)$, and a nonmembership AAD, $\hat{\nu}_{A@K}(x)$. Hence, the evaluation of the previous example can be better characterized by $\langle x, \hat{\mu}_{A@K}(x), \hat{\nu}_{A@K}(x)\rangle$, where the meaning of $\hat{\nu}_{A@K}(x)$ is analogous to the meaning of $\hat{\mu}_{A@K}(x)$, i.e., $\hat{\nu}_{A@K}(x)$ is a pair $\langle \nu_{A@K}(x), F_{\nu_{A@K}}(x)\rangle$ such that $\nu_{A@K}(x)$ represents the level to which x does not belong to the category 'wheat' and $F_{\nu_{A@K}}(x)$ is the collection of features that have been considered for quantifying the value of $\nu_{A@K}(x)$ according to K.

In the next section, we describe how to use AADs to contextualize predictions made by our novel variant of a naive Bayes classification process.

3 Explainable Naive Bayes Classification

Let F be the set of features under consideration. In *naive Bayes classification* [9,24], the probability $P(A|x)$ of an object, say x, being in a class (or category), say A, is given by

$$P(A|x) \propto P(A) \prod_{f \in x} P(f|A), \tag{1}$$

where $P(A)$ is the prior probability of x being member of A, and $P(f|A)$ is the conditional probability of a feature $f \in F$ occurring in an object x that belongs to A. This expression takes into account the "naive" assumption made in naive Bayes classification, which states that all features in x are *mutually independent*.

The actual value of $P(A|x)$ might be unknown. However, one can compute an approximation, say $\tilde{P}(A|x) = \tilde{P}(A) \prod_{f \in x} \tilde{P}(f|A)$, through a (knowledge) model obtained from a training set, say X_0. In this regard, $\tilde{P}(A)$ can be computed by means of

$$\tilde{P}(A) = \frac{|X_A|}{|X_A| + |X_{\bar{A}}|}, \tag{2}$$

where $|X_A|$ and $|X_{\bar{A}}|$ represent, in that order, the number of objects in X_0 that belong to A and the number of objects in X_0 that do not belong to A. Likewise, $\tilde{P}(f|A)$ can be computed by means of

$$\tilde{P}(f|A) = \frac{|F_A[f]|}{|F_A[f]| + |F_{\bar{A}}[f]|}, \tag{3}$$

where f denotes any of the x's features, $|F_A[f]|$ represents the number of occurrences of f in training objects that belong to A, and $|F_{\bar{A}}[f]|$ represents the number of occurrences of f in training objects that do not belong to A. In this regard, $\tilde{P}(f|A)$ can be seen as a quantification of the level to which f favors the membership of x in A.

Instead of multiplying many conditional probabilities in Eq. 1, performing the computation by summing logarithms of probabilities is preferred. Hence, the logarithm of $\tilde{P}(A|x)$ can be computed by

$$\log \tilde{P}(A|x) = \log \tilde{P}(A) + \sum_{f \in x} \log \tilde{P}(f|A). \tag{4}$$

Additionally, to avoid zeros, one can use *Laplace smoothing* [16], which adds one to each count. Thus, Eq. 4 can be rewritten as

$$\log \tilde{P}(A|x) \propto \log \frac{|X_A| + 1}{(|X_A| + |X_{\bar{A}}|) + 1} + \sum_{f \in x} \log \frac{|F_A[f]| + 1}{(|F_A[f]| + |F_{\bar{A}}[f]|) + |F_{X_0}|}, \tag{5}$$

where $|F_{X_0}|$ denotes the number of features detected in the training objects.

Given a collection of well-known classes, say $\mathcal{C}$, one can use Eq. 5 to predict the best class C for an object x by means of

$$C = \underset{A \in \mathcal{C}}{\operatorname{argmax}}(\log \tilde{P}(A|x)). \tag{6}$$

As can be noticed, Eq. 6 computes the predicted category without giving any explanation of what has been taken into account to make that prediction. For this reason, we consider that an explicit representation of the context of the evaluations made by Eq. 5 is strongly recommended. Hence, we propose our novel version of naive Bayes classification (NBC), named *explainable NBC* (XNBC), which main components: a learning process, an evaluation process and a prediction step, are described next.

3.1 Learning Process

The purpose of the learning process in XNBC is to obtain a model of what is known about a given category. Hence, a *feature-influence model* [13], which allows for the representation of the influence of features on the classification, is built with Algorithm 1. This algorithm uses a training set, X_0, and an identifier of the category, A, as input, and returns a model $K_A = \langle \hat{\mathbf{u}}_A, t_A \rangle$ as output. The model K_A is characterized by both a *directional vector* $\hat{\mathbf{u}}_A = \omega_1 \hat{\mathbf{f}}_1 + \cdots + \omega_m \hat{\mathbf{f}}_m$ and a *threshold point* t_A in a m-dimensional feature space $\mathcal{M}$, where ω_i denotes the influence of a feature f_i, which is represented by a unit vector $\hat{\mathbf{f}}_i$ in $\mathcal{M}$. As shown in Fig. 3, the model K_A can be seen as a *line* defined by $\hat{\mathbf{u}}_A$ and t_A: while the direction of $\hat{\mathbf{u}}_A$ points towards a place where the membership in A is favored, the location of t_A identifies a point where the membership in A is neither favored nor disfavored.

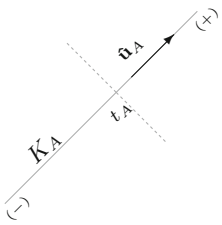

Fig. 3. Characterization of the knowledge model K_A.

To build the model, Algorithm 1 explores the objects included in the training set X_0 in order to determine the prior probability of a given object being a member of A, as well as the conditional probabilities of the features occurring in objects that belong to A. It is worth recalling that in NBC the best class for an object is considered to be the most likely. For this reason, Algorithm 1 first updates the following counters (see Lines 3–13): (i) $|X_A|$, which counts how many objects belong to the category A; (ii) $|X_{\bar{A}}|$, which counts how many objects do not belong to A; (iii) $|F_A[f]|$, which counts the occurrence of the feature f in objects that belong to A; (iv) $|F_{\bar{A}}[f]|$, which counts the occurrence of f in objects that do not belong to A. Then, the algorithm uses these counters to compute the following probabilities: (i) the prior probability $P(A)$ of an object

Algorithm 1: XNBC - Learning Process.

```
   Data: A, X_0                                   /* category, training set */
   Result: K_A                                    /* knowledge model K_A = ⟨û_A, t_A⟩ */
 1 |X_A| ← 0                                      /* number of objects that are member of A */
 2 |X_Ā| ← 0                                      /* number of objects that are nonmember of A */
 3 foreach x ∈ X_0 do
 4     if x ∈ A then                              /* if x is member */
                                                  /* ...increase the number of members */
 5         |X_A| ← |X_A| + 1
 6         foreach f ∈ x do                       /* for each f in x's features */
                                                  /* ..increase the occurrence of f in members */
 7             |F_A[f]| ← |F_A[f]| + count(f, x)
 8             F_{X_0} ← F_{X_0} ∪ {f}
 9     else                                       /* if x is nonmember */
                                                  /* ..increase the number of nonmembers */
10         |X_Ā| ← |X_Ā| + 1
11         foreach f ∈ x do                       /* for each f in x's features */
                                                  /* ..increase the occurrence of f in nonmembers */
12             |F_Ā[f]| ← |F_Ā[f]| + count(f, x)
13             F_{X_0} ← F_{X_0} ∪ {f}
   /* compute the prior probabilities                                          */
14 |X| ← |X_A| + |X_Ā|
15 P(A) ← log((|X_A| + 1)/(|X| + 1))
16 P(Ā) ← log((|X_Ā| + 1)/(|X| + 1))
   /* compute the conditional probabilities                                    */
17 foreach f ∈ F_{X_0} do
18     P(f|A) ← log((|F_A[f]| + 1)/(|F_A[f]| + |F_Ā[f]| + |F_{X_0}|))
19     P(f|Ā) ← log((|F_Ā[f]| + 1)/(|F_A[f]| + |F_Ā[f]| + |F_{X_0}|))
   /* build the feature-influence model                                        */
20 b ← P(A) − P(Ā)
21 w ← 0
22 foreach f ∈ F_{X_0} do
23     w ← w + (P(f|A) − P(f|Ā))f̂_f
24 û_A ← w/||w||
25 t_A ← −b/||w||
26 K_A ← ⟨û_A, t_A⟩
27 return K_A
```

x being in A (see Line 15); (ii) the prior probability $P(\bar{A})$ of an object x not being in A (see Line 16); (iii) the conditional probability $P(f|A)$ of a feature f occurring in an object that belongs to A (see Line 18); and (iv) the conditional probability $P(f|\bar{A})$ of a feature f occurring in an object that does not belong to A (see Line 19). These probabilities are used for computing the components of K_A, i.e., $\hat{\mathbf{u}}_A$ and t_A (see Lines 20–25). As noticed, the conditional probability of each feature is used as an indicator of its relative influence on the classification.

Algorithm 2: XNBC - Evaluation Process.

Data: x, K_A /* object, knowledge model $K_A = \langle \hat{\mathbf{u}}_A, t_A \rangle$ */
Result: $\langle x, \hat{\mu}_A(x), \hat{\nu}_A(x) \rangle$ /* augmented IFS element */
$\tilde{F}_{\mu_A}(x) \leftarrow \{\}$ /* pro-membership x's features */
$\tilde{F}_{\nu_A}(x) \leftarrow \{\}$ /* pro-nonmembership x's features */
$\tilde{\mu}_A(x) \leftarrow 0$ /* pro-membership x's score */
$\tilde{\nu}_A(x) \leftarrow 0$ /* pro-nonmembership x's score */
if $t_A < 0$ **then** /* a negative threshold favors the score */
  $\tilde{\mu}_A(x) \leftarrow \tilde{\mu}_A(x) + \text{abs}(t_A)$ /* increase the positive score of x */
else /* a positive threshold disfavors the score */
  $\tilde{\nu}_A(x) \leftarrow \tilde{\nu}_A(x) + t_A$ /* increase the negative score of x */
/* recall that $\hat{\mathbf{u}}_A = \sum_{f \in F_{X_0}} \omega_f \hat{\mathbf{f}}_f$ */
foreach $f \in x$ **do** /* for each f in x's features */
  $s_f \leftarrow \text{count}(f, x) * \omega_f$ /* compute f's influence */
  if $s_f > 0$ **then** /* if f is in favor of $x \in A$ */
    $\tilde{\mu}_A(x) \leftarrow \tilde{\mu}_A(x) + s_f$ /* increase x's positive score */
    $\tilde{F}_{\mu_A}(x) \leftarrow \tilde{F}_{\mu_A}(x) \cup \{\langle f, s_f \rangle\}$ /* and record f's influence */
  else /* f is against $x \in A$ */
    $\tilde{\nu}_A(x) \leftarrow \tilde{\nu}_A(x) + \text{abs}(s_f)$ /* increase x's negative score */
    $\tilde{F}_{\nu_A}(x) \leftarrow \tilde{F}_{\nu_A}(x) \cup \{\langle f, \text{abs}(s_f) \rangle\}$ /* and record f's influence */
/* handle the consistency condition $0 \leq \mu_A(x) + \nu_A(x) \leq 1$ */
$\text{maxLevel} \leftarrow \max(1, \tilde{\mu}_A(x) + \tilde{\nu}_A(x))$
foreach $\langle f, s_f \rangle \in \tilde{F}_{\mu_A}(x)$ **do**
  $F_{\mu_A}(x) \leftarrow F_{\mu_A}(x) \cup \{\langle f, (s_f / \text{maxLevel}) \rangle\}$
foreach $\langle f, s_f \rangle \in \tilde{F}_{\nu_A}(x)$ **do**
  $F_{\nu_A}(x) \leftarrow F_{\nu_A}(x) \cup \{\langle f, (s_f / \text{maxLevel}) \rangle\}$
$\mu_A(x) \leftarrow \tilde{\mu}_A(x) / \text{maxLevel}$
$\nu_A(x) \leftarrow \tilde{\nu}_A(x) / \text{maxLevel}$
/* finally, build the augmented IFS element */
$\hat{\mu}_A(x) \leftarrow \langle \mu_A(x), F_{\mu_A}(x) \rangle$
$\hat{\nu}_A(x) \leftarrow \langle \nu_A(x), F_{\nu_A}(x) \rangle$
return $\langle x, \hat{\mu}_A(x), \hat{\nu}_A(x) \rangle$

3.2 Evaluation Process

The purpose of the evaluation process is to obtain a contextualized evaluation of the membership of a given object in a given category. The steps of this process are described in Algorithm 2. This algorithm uses an object x and the knowledge model $K_A = \langle \hat{\mathbf{u}}_A, t_A \rangle$ for a category A as input, and builds an augmented IFS element[1] $\langle x, \hat{\mu}_A(x), \hat{\nu}_A(x) \rangle$ representing a contextualized evaluation that is returned as output.

[1] To be consistent with the notation introduced in Sect. 2, we should write $\langle x, \hat{\mu}_{A@X_0}(x), \hat{\nu}_{A@X_0}(x) \rangle$. However, for the sake of readability, we use this simplified notation hereafter.

To build a contextualized evaluation, Algorithm 2 computes both a positive score $\tilde{\mu}_A(x)$ and a negative score $\tilde{\nu}_A(x)$ of x being in category A based on the threshold point t_A and the influence of the features in the directional vector $\hat{\mathbf{u}}_A$. The positive score is increased in two cases: if t_A is negative (see Line 6); and if the influence of a feature is positive (see Line 12). Likewise, the negative score is increased in two cases: if t_A is positive (see Line 8); and if the influence of a feature is negative (see Line 15). While the conditions that arise when a positive score is increased are recorded in $\tilde{F}_{\mu_A}(x)$ (see Line 13), the conditions that arise when a negative score is increased are recorded in $\tilde{F}_{\nu_A}(x)$ (see Line 16).

The consistency condition of an IFS element, i.e., $0 \leq \mu_A(x) + \nu_A(x) \leq 1$, is guaranteed by Algorithm 2 in Lines 17–23. After this, the algorithm records the components of the augmented IFS element in Lines 24–25.

3.3 Predicting the Best Class(es)

To predict the best class $C \in \mathcal{C}$ for an object x, an XAI system (see Sect. 1) can first use Algorithm 1 for building a knowledge model for each class in $\mathcal{C}$. Then, that system can use Algorithm 2 to obtain the contextualized evaluation of the membership of x in each class using these models. After that, the system can use the buoyancy of those contextualized evaluations (see Sect. 2) to sort them in descending order. Then, the system can, say, list the top-k of the contextualized evaluations so that a user can be offered the k best classes with the best context. For each class an augmented IFS element, expressing the context of the evaluation of x belonging to the class or not, is provided. Together these explain to users why x has been classified in this way. Hence, with XNBC users and applications have extra information for giving preference to those classes with the best credible justification.

4 Illustrative Example

In this section, we present an example where our novel version of naive Bayes classification is used for predicting the classes of newswire articles. In this example, the *Reuters-21578* collection [8], which consists of 21578 newswire articles provided by Reuters, Ltd, has been used. Specifically, we made use of the articles established in the *"modified Apte split"* (ModApte) of this collection.

To use Algorithm 1, each article had to be modeled as a feature-influence vector whose components are the words in the article. Hence, each article was first split into words using separators such as commas or blank-spaces. Then, *stop words*, i.e., words like prepositions or conjunctions that have a negligible impact on the classification [11], were removed from the previous list of words. Additionally, words having a common stem were tokenized using the *Porter Stemming Algorithm* [18]. After that, Algorithm 1 was used with the feature-influence vectors corresponding to the 9603 articles included in the training set of the ModApte split for building a knowledge model for each of the following categories: earn, acq, money-fx, grain, crude, trade, interest, ship, wheat, corn.

For the sake of illustration in this paper we consider one article from the test set of the ModApte split, namely the newswire article identified by 14841. Algorithm 2 was used with the resulting knowledge models to evaluate the membership of this article to each of the aforementioned categories. This means that, augmented IFS elements like $\langle x, \hat{\mu}_{earn}(x), \hat{\nu}_{earn}(x)\rangle$ or $\langle x, \hat{\mu}_{grain}(x), \hat{\nu}_{grain}(x)\rangle$ were obtained as output – here x represents the article identified by 14841.

The resulting augmented IFS elements were used for building visual representations like the ones depicted in Fig. 4. For instance, $\mu_{grain}(x)$ and $\nu_{grain}(x)$, which are parts of $\hat{\mu}_{grain}(x)$ and $\hat{\nu}_{grain}(x)$ respectively, were used for computing the buoyancy $\rho_{grain}(x) = 0.60$ of the article in category 'grain' (see Fig. 4(a)). Analogously, the positive influence of the word 'wheat' on the membership of this article in category 'grain', namely $\langle$'wheat', $0.15\rangle \in F_{\mu_{grain}}(x)$, was used for setting both the size and the typographical style of this word. Herein, while the size of the word denotes the influence of this word on the classification, the typographical style denotes whether this influence is *positive* or negative.

Food *Department* *officials* said the *U.S.* *Department* of *Agriculture* approved the *Continental* *Grain* Co sale of 52,500 *tonnes* of *soft* *wheat* at 89 *U.S.* Dlrs a *tonne* C and F from Pacific *Northwest* to *Colombo.* They said the *shipment* was for April 8 to 20 *delivery.* REUTER

(a) $\rho_{grain}(x) = 0.60$

Food *Department* officials said the *U.S.* *Department* of *Agriculture* approved the Continental *Grain* Co sale of 52,500 *tonnes* of *soft* *wheat* at 89 *U.S.* Dlrs a *tonne* C and F from Pacific *Northwest* to *Colombo.* They said the *shipment* was for April 8 to 20 *delivery.* REUTER

(b) $\rho_{wheat}(x) = 0.42$

Food *Department* officials said the *U.S.* *Department* of *Agriculture* approved the *Continental* *Grain* Co sale of 52,500 *tonnes* of *soft* *wheat* at 89 *U.S.* Dlrs a *tonne* C and F from Pacific *Northwest* to *Colombo.* They said the *shipment* was for April 8 to 20 *delivery.* REUTER

(c) $\rho_{corn}(x) = 0.23$

Food Department *officials* said the U.S. Department of Agriculture approved the *Continental* *Grain* Co sale of 52,500 *tonnes* of *soft* *wheat* at 89 U.S. Dlrs a *tonne* C and F from Pacific *Northwest* to *Colombo.* They said the *shipment* was for April 8 to 20 delivery. REUTER

(d) $\rho_{ship}(x) = -0.47$

Fig. 4. The four best evaluated categories for a newswire article x.

Those augmented IFS elements were also used for building explanations like the following: *"While words like 'Dlrs', 'April' or 'Pacific' suggest that article* 14841 *does not belong to category 'grain' with a computed overall grade of* 0.20*, words like 'grain', 'wheat' or 'tonnes' suggest that the article belongs to the category with a computed overall grade of* 0.80*. These results indicate that article* 14841 *should be considered member of category 'grain' up to a* 0.60*-level."* Notice that this explanation indicates not only the level to which this article belongs to the category 'grain' but also provides practical information about what words

(features) have been focused on during the evaluation. We foresee that this kind of explanation can help, say, an AI practitioner to improve the knowledge model used for the evaluation. For instance, if an AI practitioner considers that 'Dlrs' and 'April' are irrelevant to the evaluation, he/she might exclude these words from the list that is used during the learning process. Notice also that only the six most influential words (three with positive influence and three with negative influence) have been included in the explanation in order to keep it simple and interpretable. A future work will reveal how this simplification could be used for improving knowledge models that result from training sets having imperfect or scarce data.

Regarding the prediction of the best category (or categories) for article 14841, the contextualized evaluations were first sorted in descending order according to the computed buoyancy. After that, the four best evaluated categories (see Fig. 4) were presented as the most optimistic predictions. As noticed, these predictions reuse the context of the evaluations and, thus, they can be easily interpreted. Hence, a user can choose the category which prediction has the most adequate justification according to his/her perspective. In this regard, experimental studies about the interpretability and usability of such predictions are considered and highly suggested.

5 Related Work

Methods aiming to produce a set of rules that explain predictions can be found in the literature. For instance, a Bayesian method for learning rules that provide explanations of the predictions according to prior parameters fixed by a user is proposed in [22]. Another example is the method proposed in [7] for building Bayesian rules that discretize a high-dimensional feature space into a series of interpretable decision statements. In the framework of fuzzy set theory, an example is the variant of the neuro-fuzzy classification method presented in [17]. This variant tries to produce a small set of interpretable fuzzy rules for the diagnosis of patients.

A comprehensive survey of methods proposed for explaining computer predictions can be found in [4]. This survey has identified two main approaches of the works found in the literature: one trying to describe how 'black box' machine learning approaches work, and the other trying to explain the result of such approaches without knowing the details on how these work. In the first approach, the goal is to make "transparent classifiers" by training interpretable models that can be used for yielding satisfactory explanations. In the second approach, the purpose is to understand the reasons for the classification or how a model behaves by, say, changing one or more inputs. In this regard, while our novel XNBC can be considered to belong to the works following the first approach, the explanation technique proposed in [19] is an example of the second approach. It is worth mentioning that techniques based on the second approach try to explain only the reasons for a specific prediction. In contrast, techniques like XNBC try to explain what has been relevant to the knowledge model and is applicable for all the possible predictions.

Contributions proposed by the fuzzy logic community for explaining computer predictions are analyzed in [1]. This analysis suggests that efforts made by the non-fuzzy community and by the fuzzy logic community can be linked to solve problems related to the interpretability of computer predictions.

6 Conclusions

In this paper, we have proposed a novel variant of a naive Bayes classification (NBC) process that produces contextualized predictions. The novel NBC process, named *explainable NBC* or XNBC, consists of a learning process, an evaluation process and a prediction step: while the purpose of the first is to obtain a model of what is known about a particular class, the purpose of the second is to obtain contextualized evaluations of the level to which other objects belong to that class, these evaluations can then be used in the third for offering users the k best classes with the best context.

The learning process looks into the objects included into a training collection to build a knowledge model in which the influence of the features on the contextualized evaluations is represented. In this process, the influence of a feature is determined by the conditional probability of the feature occurring in objects that belong to the analyzed class.

The evaluation process uses such a knowledge model as input to quantify the influence of the features on the classification of other objects. *Augmented appraisal degrees* (AADs), which are mathematical representations of the context of experienced-based evaluations, are used for handling the evaluations performed during this process. Hence, the evaluation process produces contextualized evaluations that put the forthcoming predictions in context.

In the prediction step, the k best classes corresponding to the top-k of the resulting contextualized evaluations are presented in such a way that users have additional information for giving preference to the class(es) with the best credible justification.

By means of an example in which the categories of newswire articles are predicted, we have illustrated how the proposed XNBC process can produce contextualized predictions. We have also explained how those contextualized predictions can help a user to decide which prediction is the most appropriate according to his/her perspective and, thus, make an informed (classification) decision. In spite of that, further study is needed to demonstrate the interpretability and usability of such contextualized predictions.

References

1. Alonso, J.M., Castiello, C., Mencar, C.: A bibliometric analysis of the explainable artificial intelligence research field. In: Medina, J., et al. (eds.) IPMU 2018. CCIS, vol. 853, pp. 3–15. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-91473-2_1

2. Atanassov, K.T.: Intuitionistic fuzzy sets. Fuzzy Sets Syst. **20**(1), 87–96 (1986). https://doi.org/10.1016/S0165-0114(86)80034-3
3. Atanassov, K.T.: On Intuitionistic Fuzzy Sets Theory. Studies in Fuzziness and Soft Computing, vol. 283. Springer, Heidelberg (2012). https://doi.org/10.1007/978-3-642-29127-2
4. Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., Pedreschi, D.: A survey of methods for explaining black box models. ACM Comput. Surv. **51**(5) (2018). https://doi.org/10.1145/3236009
5. Gunning, D.: Explainable Artificial Intelligence (XAI) (2017). www.darpa.mil/attachments/XAIIndustryDay_Final.pptx
6. Kaminski, M.E.: The right to explanation, explained. Berkeley Tech. LJ 34, 189 (2019). https://doi.org/10.15779/Z38TD9N83H
7. Letham, B., Rudin, C., McCormick, T.H., Madigan, D.: Interpretable classifiers using rules and Bayesian analysis: building a better stroke prediction model. Ann. Appl. Stat. **9**(3), 1350–1371 (2015). https://doi.org/10.1214/15-AOAS848
8. Lewis, D.D.: Reuters-21578 Text Categorization Collection. http://kdd.ics.uci.edu/databases/reuters21578/reuters21578.html
9. Lewis, D.D.: Naive (Bayes) at forty: the independence assumption in information retrieval. In: Nédellec, C., Rouveirol, C. (eds.) ECML 1998. LNCS, vol. 1398, pp. 4–15. Springer, Heidelberg (1998). https://doi.org/10.1007/BFb0026666
10. Lewis, D.D., Jones, K.S.: Natural language processing for information retrieval. Commun. ACM **39**(1), 92–101 (1996). https://doi.org/10.1145/234173.234210
11. Liu, B.: Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data. Springer, Heidelberg (2007). https://doi.org/10.1007/978-3-642-19460-3
12. Loor, M., De Tré, G.: On the need for augmented appraisal degrees to handle experience-based evaluations. Appl. Soft Comput. **54**, 284–295 (2017). https://doi.org/10.1016/j.asoc.2017.01.009
13. Loor, M., De Tré, G.: Identifying and properly handling context in crowdsourcing. Appl. Soft Comput. **73**, 203–214 (2018). https://doi.org/10.1016/j.asoc.2018.04.062
14. Loor, M., De Tré, G.: Explaining computer predictions with augmented appraisal degrees. In: 2019 Conference of the International Fuzzy Systems Association and the European Society for Fuzzy Logic and Technology, EUSFLAT 2019. Atlantis Press, August 2019. https://doi.org/10.2991/eusflat-19.2019.24
15. Loor, M., Tapia-Rosero, A., De Tré, G.: Usability of concordance indices in FAST-GDM problems. In: Proceedings of the 10th International Joint Conference on Computational Intelligence, IJCCI 2018, pp. 67–78 (2018). https://doi.org/10.5220/0006956500670078
16. Manning, C.D., Raghavan, P., Schütze, H.: Introduction to Information Retrieval. Cambridge University Press, Cambridge (2008)
17. Nauck, D., Kruse, R.: Obtaining interpretable fuzzy classification rules from medical data. Artif. Intell. Med. **16**(2), 149–169 (1999). https://doi.org/10.1016/S0933-3657(98)00070-0. Fuzzy Diagnosis
18. Porter, M.F., et al.: An algorithm for suffix stripping. Program **14**(3), 130–137 (1980). https://doi.org/10.1108/00330330610681286
19. Ribeiro, M.T., Singh, S., Guestrin, C.: "Why Should I Trust You?": Explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD 2016, pp. 1135–1144. ACM, New York (2016). https://doi.org/10.1145/2939672.2939778
20. Vapnik, V.N.: The Nature of Statistical Learning Theory. Springer, New York (1995). https://doi.org/10.1007/978-1-4757-2440-0

21. Vapnik, V.N., Vapnik, V.: Statistical Learning Theory, vol. 1. Wiley, New York (1998)
22. Wang, T., Rudin, C., Velez-Doshi, F., Liu, Y., Klampfl, E., MacNeille, P.: Bayesian rule sets for interpretable classification. In: 2016 IEEE 16th International Conference on Data Mining (ICDM), pp. 1269–1274, December 2016. https://doi.org/10.1109/ICDM.2016.0171
23. Zadeh, L.: Fuzzy sets. Information and control **8**(3), 338–353 (1965). https://doi.org/10.1016/S0019-9958(65)90241-X
24. Zhang, H.: The optimality of Naive Bayes. In: Proceedings of the Seventeenth International Florida Artificial Intelligence Research Society Conference, pp. 562–567. The AAAI Press, Menlo Park (2004). https://aaai.org/Papers/FLAIRS/2004/Flairs04-097.pdf