STATISTICS, OPTIMIZATION AND INFORMATION COMPUTING
Stat., Optim. Inf. Comput., Vol. **10**, February 2022, pp 225–235.
Published online in International Academic Press (www.IAPress.org)

The Optimal Inference Rules Selection for Unstructured Data Multi-Classification

Mariem Bounabi [1,*], Karim EL Moutaouakil [2], Khalid Satori [1]

[1] *LISAC, FSDM, Universite Sidi Mohamed Ben Abdellah, Fes, Morocco*
[2] *Engineering Sciences Laboratory, Multidisciplinary Faculty of Taza, Sidi Mohamed Ben Abdallah University, Taza, Morocco*

Abstract The Fuzzy Inference System (FIS) is frequently utilized in a variety of Text Mining applications. In the text processing domains, where the amount of the processed data is vast, inserting manual rules for FIS remains a real issue, especially in the text processing domains, where the size of the processed databases is enormous. Therefore, an automated and optimal inference rules (IR) selection strengthens the FIS process. In this work, we propose to apply the FP-Growth as an association model algorithm and an automatic way to identify IR for fuzzy text vectorization. Once the fuzzy vectors are generated, we call the selection variables algorithms, e.g., Info Gain and Relief, to reduce the given descriptor dimensionality. To test the new descriptor performance, we propose multi-classes text classification systems using several machine learning algorithms. Applying benchmarked databases, the new technique to produce Fuzzy descriptors achieves a significant gain in time, precision rules, and weighting quality. Moreover, comparing the classification systems, the accuracy is improved by 10% comparing with other approaches.

Keywords FIS, FTF-IDF, FP-Growth, Text Mining, ML classifiers

AMS 2010 subject classifications 62G86

DOI: 10.19139/soic-2310-5070-1131

1. Introduction

The automatization of expert systems is a challenge in several areas [1] [2]. The main aim is to show, monitor, and provide relevant information utilizing fast and intelligent technologies, particularly in the artificial intelligence context applied to textual data. For that reason, experts in the development of several beneficial systems, notably in the text mining field, apply fuzzy inference systems[3].

One of the FIS use in the Text Mining field is the technique of weighting features (FTF-IDF) [4], where we use fuzzy reasoning to extract the term frequency-inverse term frequency (TF-IDF) scores [5]. The method used by the FIS system to generate the FTF-IDF matrix descriptor [4] takes time away from the definition of the inference rules, and the purpose of the relevant regulations is hard to establish. Likewise, the large size of input data requests an automated generation of inference rules. Subsequently, the paper [6] proposes the association models [7] as a computed technique to solve the presented problem. Regularly, the given approach process are: Preprocessing of the employed dataset, nominal representation using membership function, the association model call, and a posttreatment applied on the generated rules [8].The proposed approach applies the Apriori algorithm to select valuable rules basing on a confidence score more than 90%.

*Correspondence to: Mariem Bounabi (Email: mariem.bounabi@usmba.ac.ma). Department of Computer Science, Signals, Automation and Cognitivism Laboratory (LISAC), Faculty of Sciences Dhar El Mahraz-Fes, Universite Sidi Mohamed Ben Abdellah, Fes, Morocco.

ISSN 2310-5070 (online) ISSN 2311-004X (print)

In the same way, in this paper, we propose to use the FP-Growth association model to select inference rules for a fuzzy TF-DF inference system. Several FP-Growth advantages recommend its application, e.g., more efficient for large, applied data, which is the case when processing unstructured web data. Moreover, the application of FP-Growth reduces the cost of research, and due to its structure (FP-Tree), it is more efficient than Apriori. Also, unlike Apriori, which produces candidate itemsets and tests them to keep only frequent itemsets, FP- Growth constructs frequent itemsets without generating candidates [9]. In this contribution, the FP-Growth allows producing more explicit rules, which minimize the need for post-processing as a complicated step.Additionally, the experiments prove that the new technique permits an optimal, rapid, and interoperable selection of inference rules to produce relevant Fuzzy Descriptors. As the second part of our presented contribution, we cote to generate fuzzy descriptors using the mentioned approach for several textual corpora for automatic multi-classes text classification. This level permits to test the performance of the given descriptor, where we compare several unstructured data categorization systems using:

- The selection features methods, e.g., Relief and Info Gain, reduce the descriptor's dimensionality and enhance the performance of the used classifiers.
- A set of Machin Learning (ML) classifiers, i.e., Bayes Network, SVM, Random Forest, and the Vote as a hybrid technique to combine several used ML classifiers.

Consequently, we confirm the improvement of the FTF-IDF based on the used association models, comparing with the fuzzy representation using the traditional FIS on the text classification systems.

For more details, we propose to organize our paper as follows: The related works are present in the next section, accompanied by the proposed approach part, where we quote the set of selection rules process and the reason for our contribution, accompanied by the adopted unstructured classification architecture. Next, we define the used Machin Learning (ML) tools. To show the compared multi-classes text classification systems performances, we present the experimentation and results in the fourth section before the conclusion.

2. Related Work

The Boolean logic (BL) is the first reasoning, which uses the classical membership functions varying in the range True, False or 0, 1. Therefore, the BL applications in imprecise and uncertain domains are limited [10]. In this way, a binary representation of textual data was adopted by [11], but this approach's weaknesses, e.g., information loss, made it excluded and ineffective [11]. Later, Zadeh [12] was invented the Fuzzy Logic (FL), as a classical logic extension, where the membership functions have values in the range [0, 1]. The FL has significant attention from the researchers, whose main aim is to manipulate natural language's uncertain notions [13]. In this context, a new membership function has been defined: $\mu \cdot A : X[0, 1]$, which implies that x belongs to the fuzzy set A, with a degree of truth equal to $\mu \cdot A(x)$ and like that each information will be valued without loss. Generally, the fuzzy set theory can be applied in a wide range of domains. The information is incomplete or imprecise, e.g., classification [14], pattern recognition [15], data mining [16], information systems [17], and much other artificial intelligence [18, 19]. In this work, we deal with the fuzzy weight FTF-IDF applications used by Gupta, al [20] to define a novel fuzzy logic-based ranking function for an efficient information retrieval system. In this regard, the first part of the given scenario in [4] uses the membership function $\mu \cdot W(x)$ to determine each term's representative score in a document.

Generally, to make a fuzzy decision, it is necessary to follow the three principal phases for the Fuzzy Inference System (FIS). The first process is Fuzzification methods [21] which consists of characterizing the system's linguistic variables [22, 23]. Next, the Inference phase [24], where the inference engine is applied for condensing information of a system based on a set of rules, expressed by experts. The closing process is the defuzzification phase [12], which produces the final output using existing methods. Notably, in the fuzzy deductions or inferences step, the system uses a hand-operated manner to determine rules. consequently, for various kinds of rules, e.g., single rule inference type and several inference rules [25], it necessarily needs automatization. In our case, the

used FIS requires a set of precise and relevant inference rules to generate a robust descriptor that develops the text classification process [6]. Thus, to control the employed data, we were forced to automate making inference rules, which influences the desired results.

Since association rules are widely used with fuzzy logic in different contexts, e.g., [26, 27], and Yi-Chung Hu et al. [28], how introduced a classification approach uses data mining techniques to obtain a better set of fuzzy rules for categorization tasks. Several Algorithms, for extracting frequent items (association rules), are proposed in the literature such as Apriori algorithm [29], Close [30], OCD [31], Partition [32], DIC Dynamic Items and Counting [33], Tertius algorithm [34]. The Apriori and Filter associations [6] are usually used because of their simplicity [6]. In 2000, FP-Growth [7] was discovered with several advantages. Unlike Apriori, the FP-Growth method transforms the problem of finding the longest frequent itemset by finding the smallest and its concatenation with the corresponding suffix, which permit to reduce the cost of the research. Based on the cited works, we propose a new system that calculates the fuzzy terms matrix scores (FTF-IDF scores), using association model for the inference rules process, where we chose FP-Growth, because of its structure (FP-Tree), which makes it more efficient than Apriori and other association models.

Due to the influence of the term weighting techniques on the unstructured data categorization [35], we propose to visualize the generated FTF-IDF performance on the text classification systems accuracies. In the next section we describe several used methods to make text classification with the new approach.

3. The Proposed Approach

3.1. Main observations and motivations

Before detailing the adopted system components, we want to defend the reasons for practicing this approach. As it was mentioned, in our article [6], the fuzzy representation FTF-IDF is highly recommended, in the literature, for vector representation, and several works like [4, 20] have proved that the FTF-IDF influences the quality of the desired systems. For that, we suggest improving the used FIS to generate FTF-IDF weights by automating the inference rule (IR) process, where the association rules have been incorporated. In [6] we have proved the positive impact of automatic inference rules proposal on the quality of the fuzzy representation and the unstructured document classification systems. Hence, differently to popular approaches, e.g., [36], we suggest changing the Apriori algorithm by the FP-Growth as an association model in the automatization IR approach. The FP-Growth is more efficient for large sets of data; it allows producing more explicit rules, which minimize the need for post-processing as a complicated step. Additionally, the new technique permits an optimal, rapid, and interoperable selection of inference rules to produce Fuzzy Descriptors (FD). To show the FD effectiveness, we propose to test with other sensitive ML classifiers than the Bayesian classifiers, to classify unstructured corpus. Subsequently, we use the given descriptor to test with different categorization systems using as classifiers: SVM, Bayesian network, and Random Forest combined, by the vote technique to present an efficient unstructured data classification system.

3.2. The adopted Fuzzy TF-IDF approach

To determine the FTF-IDF weight, Figure 1 shows that it is necessary to follow the fuzzy inference system process.

The fuzzification phase is the first step, where the crisp or real inputs are the term frequency (TF), Inverse Document Frequency (IDF), and N as the length of the document. Also, the TF-IDF components are assigned by membership degrees using the membership functions tool. Regularly, the membership function curves are triangular, trapezoidal, Gaussian, sigmoid, and polynomial forms. As shown in [4], the used model in this work is the triangular membership function, and any other choice has an identical impact.

Furthermore, linguistic inputs and output are the set of corresponding values designed in [4]. After determining the linguistic variables, FIS's second process converts the fuzzy input to the fuzzy output using the fuzzy rules. Usually, the rules are a set of linguistic statements that follow human expert knowledge and empirical rules. This

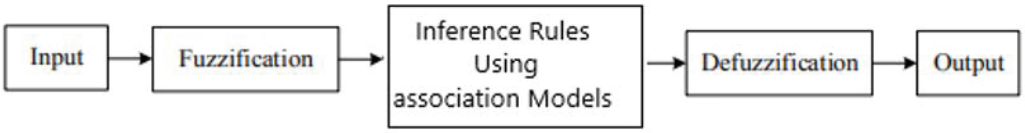

Figure 1. The adopted Fuzzy inference system process for FTF-IDF.

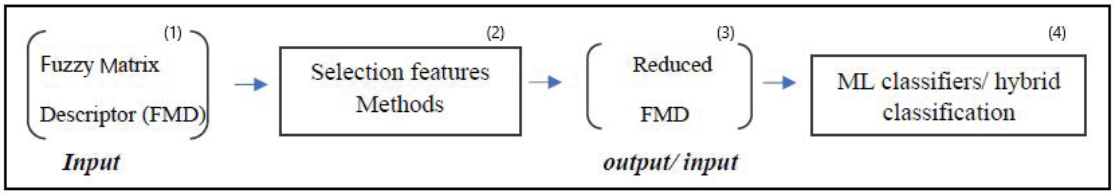

Figure 2. the adopted architecture for unstructured data classification.

step must be revisited to automatize the process, mainly when we utilize a big data size. In this approach, we use the IF-THEN type of fuzzy rules generated automatically by association models. Moreover, to realize the proposed rules automatization, a set of processes was adopted:

1. The first step is the data pre-processing, which has an essential effect on several systems. For that, we use mathematical expressions to model the process, where we consider the set of data as $D = \{x_1 x_k x_N\} \subset R^n$ by decomposing $\forall i \in [1, n]$ the interval $[min^k {x_i}^k, max^k {x_i}^k]$, and we define the appropriate membership functions.
2. Secondly, the use of Membership functions, for every input and output, permit the Nominal representation of the data set. Consequently, we determine which fuzzy membership functions, for each simple attribute X, that fall within, and the component of d are substituted with these fuzzy sets using the logical operators AND, and OR.
3. In the third stage, we call the association models to select the best rules basing on the confidence level. In the third stage, we call the association models to select the best rules basing on the confidence level. Several association models were proposed in the literature, and in this work, we use the FP-Growth model.
4. Finally, Post-treatment, as the last process in our approach, is applied to find further relevant rules. A set of operations, in the Post-treatment stage, were proposed in [6].

In our case, the implied approach permits us to make Fuzzy weights for textual corpus terms. Generally, the term weighting technique has a direct impact on the text categorization system. For that, we test the quality of the given representation by FD inclusion into several classification systems. Hence, once the matrix descriptor (Fuzzy Matrix Descriptor (1)) was generated by the FIS, we follow the text categorization architecture presented in figure 2:

- The first proposed step is the feature selection using as inputs the FMD for several attributes selection methods. In this work, we compare relief and info gain performance inside the text classification systems.
- The use of feature selection methods reduces the descriptor dimensionality and enhances the classifier's accuracies. In this step we obtained the reduced FMD (3).
- After that, we call the ML classifiers (4), i.e., SVM, Bayesian networks, and random forest, to test the descriptor performance. To improve the classification system, we propose to combine the set of used classifiers employing the vote technique as a hybrid classification. In the sections to follow, we describe several used ML tools.

4. Machine Learning Models

In this section, we present the association rules algorithms as type ML models, which permit an automatic selection of relevant rules for our approach to produce representative vectors for the used datasets. Also, once the modified FIS generates the descriptor, we call some ML classifiers like Bayesians classifiers to show the impact of the given representation on the text classification task. For that, the description of Bayesian networks is given likewise in this part.

4.1. Association Models

Generally, an association rule (A B) is usually composed of a set of items $I\{i_1, i_2, i_3, etc.\}$. A set of transactions $T\{t_1, t_2, t_3, t_4, etc.\}$ corresponds to a learning set which will be used to determine the association's rules in the following step of the model. The volume of the transaction is the number of items contained in the transaction. An important notion for a set of items is its support θ , which refers to the number of observed transactions that contain it. The support (S) and confidence (C) are the performance measures of an association rule, defined as follow:

$$\sigma(X) = Card\{t_i/X \subseteq t_i, t_i \in T\} \tag{1}$$

$$S(A \to B) = \sigma(A \cup B)/N \tag{2}$$

Where:
$Support$: The rule occurrence in the used data.
$\sigma(AUB)$: support of the items $A \cup B$;
N: the total number of transactions.
$Confidence$: measure the rule validity (percentage of examples that verify the conclusion).

$$C(A \to B) = \sigma(A \cup B)/\sigma(A) \tag{3}$$

It must be mentioned that a rule with a low support can be observed only by chance. For a set of a transaction, we can generate rules by founding all association rules with support $\geq$ min support and confidence $\geq$ min confidence, where min support and min confidence are thresholds for support and confidence [6].

4.2. FP-Growth

FP-Growth is a model applied to generate rules whose components of the rules are linked by consequence relation. The given algorithm [35] constructs frequent item sets without generating candidates, unlike Apriori [29], which generates candidate item sets and tests them to keep only frequent item sets. Moreover, the FP-Growth uses the compact tree structure (FP-tree) to compress a big database. Generally, the FP-growth process in six main steps:

1. Browse the database the first time to get the frequencies of an itemset.
2. Sort the items and in decreasing order according to their frequency.
3. A second route for the construction of the FP Tree.
4. Build a conditional pattern base for each node.
5. Build conditional FP-tree for each node.
6. Generate the most frequent patterns. The performance studies show that FP-Growth is faster than Apriori and then the tree-projection. Thus, due to the structure (FP-Tree) used by the FP-Growth, makes it more efficient.

4.3. Select attributes Methods

One of the central problems of machine learning is identifying a representative set of features from which to construct an effective classification model. In this work, we compare three methods to improve the quality of the classification.

- **Relief**: is an algorithm that takes a filter-method approach to feature selection. Relief calculates a feature score for each feature which can then be applied to rank and select top-scoring features for feature selection [37].
- **Info gain**: uses to greedily select points into the active set, by measuring how much information a feature gives us about the class. To use this classification tools, we process in four phases [38].

4.4. Classification Tools

Most classification methods perform model construction based on feature vectors by converting terms per document into a representation, which is comprehensible to a learner by choosing the suitable Data Mining algorithm according to the user goals. Here we have the types of classification algorithms in Machine Learning:

- **Bayesian networks (BN)** [5] Are acyclic probabilistic graphs, who are the nodes are random variables, the structure of the network defines their conditional dependencies. To use this classification tools, we process in four phases:
  1. Modeling the problem n terms of a set of random variables $X = \{X_1, ..., X_n\}$
  2. Choosing an adequate architecture of the network,
  3. Constructing the conditional probabilities matrix of node i, knowing the state of its parents:
  $$\theta = P(X_i/P_a(X_i)) \tag{4}$$
  4. Meddling with a given request's answer. Such a task is based on the joint distribution:
  $$P(X_1, X_2, ..., X_n) = \prod P(X_i/P_a(X_i))) \tag{5}$$

  Different algorithms, for learning the network structure, were proposed in the literature such as Hill Climber, Tabu Search, and K2; in our case, the K2 is the used structure to learn the network [39].
- **Naive Bayes (NB)** [5] is a set of supervised learning algorithms based on applying Bayes theorem with the "naive" assumption of independence between every pair of features. Basing on this naive assumption, equation 4 is simplified to:
  $$P(y/x_1, x_2, , x_n) = (P(y) \prod P(x_i/y))/P(x_1, x_2, , x_n) \tag{6}$$

  Since $P(x_1, x_2, , x_n)$ is constant given the input,we can use the provided classification rule:
  $$Y = Arg.max\ P(y) \prod P(x_i/y) \tag{7}$$
- **Naive Bayes updateable (BayUpdat)** [40] is the updateable version of Naive Bayes. This classifier will use a default precision of 0.1 for numeric attributes when build classifier is called with zero training instances.
- **Random Forest** [5] The random forests include an ensemble learning method for classification, regression, and other tasks. This kind of tool operates by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees; By constructing a multitude of decision trees, this approach overcomes the overfitting problem.
- **Support Vector Machine** [41]: Are supervised learning models with associated learning algorithms that analyze data used for classification and regression analysis. The main goals of this algorithm are to find a hyperplane in an N-dimensional space of the number of features that distinctly classify the data points.
- **Hybrid ML models** [41]: To reach the best performances of the tested classifiers, we use the hybrid machine learning models using some variants of majority voting systems by varying the combiner function. The hybrid techniques are often, used to improve the performance of unstable or weak classifiers. To combine classifiers, the voting method finds what is the class output of each classifier and count its output, as a vote, for a class, by assigning the input pattern to the class with the majority vote. Several combiner functions were proposed, which contained the sum, product, max, min, average and median functions.

Table 1. BBCSport class and number of articles

Class	Number of articles
Athletics	101
Cricket	124
Football	256
Rugby	147
Tennis	100

Table 2. BBCNEWS class and number of articles

Class	Number of articles
Business	510
Entertainment	386
Politics	417
Sport	511
Technology	401

5. Experiments & Results

The whole of the algorithms studies has been implemented with java language, which favors our comparative study. To compare different recognition systems, based on the new fuzzy representation FTF-IDF approach and classifiers, several experimentations have been conducting for all algorithms with different configurations under a compatible Dell, Intel (R) Core i5- CPU 2.50 GHz, and 4 GB of RAM.

5.1. Experimentation setup

5.1.1. Datasets We use as corpus the BBC News and BBC Sport data news [42] classified in five predictive classes; To improve the performance of the used classifiers we use some attribute selection methods, widely used in the text classification.

- **BBCSport database:** is composed of 737 documents organized in five class labels mentioned on the table 1
- **BBC News database:** the corpus consists of 2225 documents classified in five predictive classes as described in the table 2

5.1.2. Pre-processing Text preprocessing or cleaning is an essential preparatory stage in news categorization that decreases the needed space and improves classification efficiency. Most of the time, the dataset is unstructured and contains a mix of relevant and worthless information. Stop words, punctuation, special characters, unrelated phrases, quotes, and dates are unnecessary information that adds no predictive value to the classifier/model. They take up space and have the potential to skew the ML model. As a result, before extracting any feature from the raw dataset, a cleaning step should be carried out to reduce the distortions brought into the model [13]. Several procedures were taken in this study to preprocess the news text:

1. Transforming text in the same letter size.
2. Removing Characters such as ?, !, ; and . .
3. Filtering Stop Words.
4. Using stemmers to save radical of terms.

Table 3. Measures performance for a classifier

Measure	Formula
Recall	$TP/TP+FP$
Precision	$TP/TP+FN$
F-Measure	$(2*Precision*Recall)/(Precision+Recall)$
Accuracy	$(TP+TN)/(TP+FP+TN+FN)$
AUC [43]	$(Recall+sensibility)/2$
Specificity	$TN/TN+FP$

To improve the performance of the used classifiers we use some attribute selection methods, widely used in the text classification.

5.1.3. The Classification parameters Each classifier has its parameters, which allow the improvement of the tasks according to the problem.

Support Vector Machine: The kernel type characterizes the SVM model and can change the classification results, in our case we use the polynomial kernel, which is a kernel function largely used with the SVM model and it permits us to learn any Model types [41].

Bayesian Classifier: For the Bayesian approach, one of the main problems is to find the optimal network structure, thus, we tested different algorithms for learning the network structure (Hill Climber, Tabu Search, K2, and Tree Augmented Naive). We present, in the next section, the results of the search K2 network structure, which gives the best score.

Vote technique: To improve the system performance, we combine all these classifiers using the hybrid classifiers system with different combiner functions such as average, product, and majority. The majority voting achieves satisfactory results compared to the other combiner functions [41]. For that, we use the majority vote scores to analyze the performance of the proposed systems.

5.2. Performance Measures

The effectiveness of the classifier can be evaluated by a set of measures [6]. Table 3 summarizes a set of the used measures, which evaluate the reliability of the classification mechanism.

The given musers are based on the: True Positive (TP), False Positive (FP), False Negative (FN), True Negative (TN) metrics explained on [5].

5.3. Results & Discussion

5.3.1. BBC Sport Dataset Using the BBCSport, table 4 shows a global comparison of different used representation, tested with the SVM, Bayes networks, random forest classifiers combined with the selected feature method, i.e., Relief and Info gain to ameliorate the classification performance. In this regard, we use the accuracy as performance measure to compare the set of systems.

Effectively, the proposed method, which is the ameliorate FTF-IDF with Relief or info gain, achieves satisfactory classification results, where the classification of the BBC Sport news (using: ML classifiers and the vote technique), as shown in Table 2, note significant accuracy progress comparing to simple FTF-IDF results.

The given accuracy = 98%, in table 4, and the presented results for the paper [6] prove that the automatization of rules to produce the FTF-IDF weight has an excellent impact on the text representation and the supervised decision.

Table 4. The results of the systems based on new FTF-IDF descriptor compared with the traditional FTF-IDF using as select Features methods Relief and Info gain, a set of classifiers, and the BBCSport data.

	FTF-IDF+ info gain				The new FTF-IDF +info gain				FTF-IDF + Relief				The New FTF-IDF + Relief			
	Recall% (R%)	Precision% (P%)	Accuracy% (A%)	AUC%	R%	P%	A%	AUC%	R%	P%	A%	AUC%	R%	P%	A%	AUC%
Bayes Net	95.2	95	95	97.8	97.1	97	97	99.5	95	95	95	97.5	97	97	97	98.2
Bayes Naive	94	94	94	98.7	96	96	96	99.5	95	95	95	96.9	97	97	97	98.2
Bays update	98	98	98	98.9	98.3	98.3	98.3	99	98	98	98	99	98.2	98	98	98.9
SVM	93	93	92.4	95	95.5	95.5	95.5	96.7	93	93	93	95.6	94.6	94.6	94.6	97
Random Forest	93	92	92.13	94	95.5	95.5	95.5	96	91.9	91	91.5	92	91	91	91	92.9
Vote	97	96.8	96.8	98.7	98	98.7	98	99	95.2	95.4	95	97.9	96.3	96.4	96	99

Table 5. The results of the systems based on new FTF-IDF descriptor compared with the traditional FTF-IDF using as select Features methods Relief and Info gain, a set of classifiers, and the BBC News data.

	FTF-IDF+ info gain				The new FTF-IDF +info gain				FTF-IDF + Relief				The New FTF-IDF + Relief			
	Recall% (R%)	Precision% (P%)	Accuracy% (A%)	AUC%	R%	P%	A%	AUC%	R%	P%	A%	AUC%	R%	P%	A%	AUC%
Bayes Net	93	93	93	96	96.1	96	96	98.5	94.6	94.6	94.6	96	95	95	95	97.9
Bayes Naive	94	94	94	97.7	97	97	97	98	94	94	94	97	97.6	97.6	97.6	98.4
Bays update	93	93	93.03	96.0	96.01	96	96	97.6	92.6	92.6	96.2	96	96	96	96	97.9
SVM	71	71.5	71.4	87.3	90	90	90	92.2	72	72	72	90	90	91	90.9	91.8
Random Forest	80	80	80	90	90.2	90.2	90	92	83	82	82.8	90	90	90	90	92
Vote	90	90	90	91.9	94	94	94	94	90	90	90	91	95.6	95.6	95	98

5.3.2. BBC News Dataset Using other databases, the BBC News data, table 5 presents the comparison of different FTF-IDF representation, tested with several ML classifiers combined with the selected feature method, i.e., Relief and Info gain to ameliorate the classification performance. Also, a set of performance measurements are employed to compare the given systems.

The proposed term weighting technique notes a great improvement for systems based on the SVM and Random Forest classifiers, as shown in table 5. The given results describe the efficacity of the proposed term scores applying FP-Growth algorithm, where the use of the FTF -IDF with SVM and Random Forest improves the accuracy by 10%. Besides, the given results prove that the vote technique enhance the SVM and Random forest accuracies using the new FTF-IDF and the selection variables methods.

Another way, which demonstrates the new representation efficiency, is the AUC scores. Hence, the given scores for the compared systems (tables 4 and 5), using the booth databases, show that the classification systems performance based on the New Fuzzy descriptors has the satisfactory quality and excellent AUC quantity. Therefore, various papers in the literature present different multi- classes unstructured data classification like the authors of [44], who show the classification results of the BBC News database. The contributors use many instances, from the BBC News, lower than that considered in this study (1490 Vs 2225 documents). The results presented are based on:

- A simple TF-IDF vectorization.
- An approach that does not use variable selection methods as an essential step for processing and classifying large datasets.

Comparing with [44] our approach deals with the high dimensionality dataset classification. Similarly, the proposed systems are interoperable. The use of the Vote classifier in our work enhances the reliability and the security of the given systems. Thus, the use of the Fuzzy descriptor with the mentioned text classification process present excellent accuracies. Generally, the performances are higher than the recognition rates given in [44].

6. Conclusion

Some approaches propose the association models as ML tools to automatically predict IR from the set of data. The purpose is to improve the fuzzy decision-making systems responses without expert intervention. In this work, we applied the advantageous FP-Growth instead of the existing association method in the literature to determine the FTF-IDF weights for textual data. The new proposition is more efficient, where the used algorithm accelerates the processing of large datasets size. Besides, it allows producing more explicit rules, which minimize the need for post-processing as a complicated step in the automatic selection approach of IR. The new technique permits an optimal, rapid, and compatible IR selection to produce Fuzzy Descriptors (FD). Including the new FD in a multi-classes text classification system presents excellent results comparing with several text vectorization approaches. The compared systems use the new FD and the reduction dimensionality algorithms with several sensitive ML classifiers, e.g., SVM, Bayes Network, Random Forest combined using the Vote. Hence, the given approach advantages are to present a high FD quality, which impacts the multi-classes text classification performance. Indeed, the provided systems show excellent accuracies, up to 97%, for several benchmarked databases. Also, compared with a group of popular weighting schemes, our approach usually gives the best recognition rates. Generally, if we test on the average dataset, our proposal encourages practitioners to control intelligent systems to apply it for even complex problems. As future works, we suggest solving the limitation of our proposal, where we will try to deal with the real-time Text Mining applications and using other unstructured databases. Also, using our approach for different decisional Big Data systems.

REFERENCES

1. Sahin, S., Tolun, M. R. & Hassanpour, R. Hybrid expert systems: A survey of current approaches and applications. Expert systems with applications 39, 4609-4617 (2012).
2. Unnikrishnan, P., Govindan, V. & Kumar, S. M. Enhanced sparse representation classifier for text classification. Expert Systems with Applications 129, 260-272 (2019).
3. Miller, T. Explanation in artificial intelligence: Insights from the social sciences. Artificial intelligence 267, 1-38 (2019).
4. Bounabi, M., El Moutaouakil, K. & Satori, K. Text classification using Fuzzy TF-IDF and Machine Learning Models in Proceedings of the 4th International Conference on Big Data and Internet of Things (2019), 1-6.
5. Bounabi, M., Moutaouakil, K. E. & Satori, K. A comparison of text classification methods using different stemming techniques. International Journal of Computer Applications in Technology 60, 298-306 (2019).
6. Bounabi, M., Moutaouakil, K. E. & Satori, K. The Automatic option of inference rules for the fuzzy TF-IDF in 2020 IEEE 2nd International Conference on Electronics, Control, Optimization and Computer Science (ICECOCS) (2020), 1-6.
7. Agrawal, R., Srikant, R., et al. Fast algorithms for mining association rules in Proc. 20th int. conf. very large data bases, VLDB 1215 (1994), 487-499.
8. Bounabi, M., El Moutaouakil, K. & Satori, K. Association models to select the best rules for fuzzy inference system. Adv. Intell. Syst. Comput 1076, 349-357 (2020).
9. Borgelt, C. An Implementation of the FP-growth Algorithm in Proceedings of the 1st international workshop on open source data mining: frequent pattern mining implementations (2005), 1-5.
10. Fox, E. A. Extending the Boolean and vector space models of information retrieval with p-norm queries and multiple concept types PhD thesis (Cornell University, 1983).
11. Goguen, J. A. (1973). "LA Zadeh. Fuzzy sets. Information and control, vol. 8 (1965), pp. 338-353. -LA Zadeh. Similarity relations and fuzzy orderings. Information sciences, vol. 3 (1971), pp. 177-200.", The Journal of Symbolic Logic, 38(4), 656-657.
12. Kazuo, T. An Introduction to Fuzzy Logic for Practical Applications, 1997.
13. Ross, T. J. Fuzzy logic with engineering applications (John Wiley & Sons, 2005).
14. Sarikh, S., Raoufi, M., Bennouna, A. & Ikken, B. Characteristic curve diagnosis based on fuzzy classification for a reliable photovoltaic fault monitoring. Sustainable Energy Technologies and Assessments 43, 100958 (2021).
15. Spiliotis, M., Iglesias, A. & Garrote, L. A multicriteria fuzzy pattern recognition approach for assessing the vulnerability to drought: Mediterranean region. Evolving Systems 12, 109-122 (2021).
16. Wu, Y., Wang, Z. & Wang, S. Human Resource Allocation Based on Fuzzy Data Mining Algorithm. Complexity 2021 (2021).
17. Wei, Y. & Wang, Q. Nested Structures Study in Dual Hesitant Fuzzy Information Systems in 2021 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE) (2021), 1-6.
18. Shaocheng, T., Jiantao, T. & Tao, W. Fuzzy adaptive control of multivariable nonlinear systems1. Fuzzy sets and systems 111, 153-167 (2000).
19. Alonso Moral, J. M., Castiello, C., Magdalena, L. & Mencar, C. in Explainable Fuzzy Systems 1-23 (Springer, 2021).
20. Gupta, Y., Saini, A. & Saxena, A. A new fuzzy logic based ranking function for efficient information retrieval system. Expert Systems with Applications 42, 1223-1234 (2015).
21. Jang, J.-S. ANFIS: adaptive-network-based fuzzy inference system. IEEE transactions on systems, man, and cybernetics 23, 665-685 (1993).

22. Hadroug, N., Hafaifa, A., Alili, B., Iratni, A. & Chen, X. Fuzzy Diagnostic Strategy Implementation for Gas Turbine Vibrations Faults Detection: Towards a Characterization of Symptom-fault Correlations. Journal of Vibration Engineering & Technologies, 1-27 (2021).
23. Kalibatiene, D. & Miliauskaite, J. A dynamic fuzzification approach for interval type-2 membership function development: case study for QoS planning. Soft Computing, 1-19 (2021).
24. Babuska, R. & Verbruggen, H. B. An overview of fuzzy modeling for control. Control Engineering Practice 4, 1593-1606 (1996).
25. Zadeh, L. A. Is there a need for fuzzy logic? Information sciences 178, 2751-2779 (2008).
26. Zheng, H., He, J., Huang, G., Zhang, Y. & Wang, H. Dynamic optimisation based fuzzy association rule mining method. International Journal of Machine Learning and Cybernetics 10, 2187-2198 (2019).
27. El-Semary, A., Edmonds, J., Gonzalez-Pino, J. & Papa, M. Applying data mining of fuzzy association rules to network intrusion detection in the Proceedings of Workshop on Information Assurance United States Military Academy (2006).
28. Hu, Y.-C., Chen, R.-S. & Tzeng, G.-H. Finding fuzzy classification rules using data mining techniques. Pattern recognition letters 24, 509-519 (2003).
29. Bathla, H. & Kathuria, K. Apriori algorithm and filtered associator in association rule mining. International Journal of Computer Science and Mobile Computing 4, 299-306 (2015).
30. Puri, S. & Singh, S. P. in Computing and Network Sustainability 227-237 (Springer, 2019).
31. Pasquier, N. Extraction de Bases pour les Règles d'Association à partir des Itemsets Fermés Fréquents in Inforsid'2000 Congress (2000), 56-77.
32. Mannila, H., Toivonen, H. & Verkamo, A. I. E cient algorithms for discovering association rules in KDD-94: AAAI workshop on Knowledge Discovery in Databases (1994), 181-192.
33. Agrawal, R. & Shafer, J. C. Parallel mining of association rules. IEEE Transactions on knowledge and Data Engineering 8, 962-969 (1996).
34. Brin, S., Motwani, R., Ullman, J. D. & Tsur, S. Dynamic itemset counting and implication rules for market basket data in Proceedings of the 1997 ACM SIGMOD international conference on Management of data (1997), 255-264.
35. Han, J., Pei, J. & Yin, Y. Mining frequent patterns without candidate generation. ACM sigmod record 29, 1-12 (2000).
36. Vetriselvi, T. & Gopalan, N. An improved key term weightage algorithm for text summarization using local context information and fuzzy graph sentence score. Journal of Ambient Intelligence and Humanized Computing 12, 4609-4618 (2021).
37. Urbanowicz, R. J., Meeker, M., La Cava, W., Olson, R. S. & Moore, J. H. Relief-based feature selection: Introduction and review. Journal of biomedical informatics 85, 189-203 (2018).
38. Blessie, E. C. & Karthikeyan, E. Sigmis: A feature selection algorithm using correlation based method. Journal of Algorithms & Computational Technology 6, 385-394 (2012).
39. Bouckaert, R. R. Bayesian network classifiers in weka for version 3-5-7. Artificial Intelligence Tools 11, 369-387 (2008).
40. McCallum, A., Nigam, K., et al. A comparison of event models for naive bayes text classification in AAAI-98 workshop on learning for text categorization 752 (1998), 41-48.
41. Aharrane, N., El Moutaouakil, K. & Satori, K. A comparison of supervised classification methods for a statistical set of features: Application: Amazigh OCR in 2015 Intelligent Systems and Computer Vision (ISCV) (2015), 1-8.
42. Greene, D. & Cunningham, P. Practical solutions to the problem of diagonal dominance in kernel document clustering in Proceedings of the 23rd international conference on Machine learning (2006), 377-384.
43. Sokolova, M., Japkowicz, N. & Szpakowicz, S. Beyond accuracy, F-score and ROC: a family of discriminant measures for performance evaluation in Australasian joint conference on artificial intelligence (2006), 1015-1021.
44. Hussain, A., Ali, G., Akhtar, F., Khand, Z. H. & Ali, A. Design and Analysis of News Category Predictor. Engineering, Technology & Applied Science Research 10, 6380-6385 (2020).