International Journal of Computer Information Systems and Industrial Management Applications
ISSN 2150-7988 Volume 3 (2011) pp. 210-217
© MIR Labs, www.mirlabs.net/ijcisim/index.html

Fuzzy Rules for Document Classification to Improve Information Retrieval

Tatiane M. Nogueira[1], Heloisa A. Camargo[2] and Solange O. Rezende[3]

[1] Institute of Mathematics and Computer Science, University of Sao Paulo
Av. do Trabalhador são-carlense, 400, Sao Carlos-SP, Brazil
{tatiane, solange}@icmc.usp.br

[2] Department of Computer Science, Federal University of Sao Carlos
Rod. Washington Luís, Sao Carlos-SP, Brazil
heloisa@dc.ufscar.br

***Abstract*: In this work, we present a method to generate, from text documents, fuzzy rules used to classify documents and to improve the information retrieval. With this method, we face the issue of dimensionality in text documents for information retrieval. We also present a comparison analysis among the method that we proposed and well-known machine learning methods for classification. The aim of our work is to develop a mechanism to reduce the high dimensionality of the attribute-value matrix obtained from the documents and, consequently, scale up the proposed classifier. Some experiments have been run using different domains in order to validate the proposed approach and compare the results with the ones obtained with the OneR, K-Nearest Neighbor classifier, C4.5, Multi-variable Naive Bayes, and SVM methods. The experiments and the obtained results showed that this is a promising approach to deal with the dimensionality problem of document for information retrieval.**

Keywords: fuzzy clustering, information retrieval, text mining, text categorization, uncertainty, imprecision.

I. Introduction

With the popularization of the Internet, a lot of on-line information is generated, and the use of systems for collection and storage of digital information by different organizations is also increased. According to Herrera-Viedma and López-Herrera in [1], the amount of information available makes necessary the development and use of effective Information Access Systems (IASs) that allow to the user the easy and flexible access to quality and relevant information.

When there is a lot of textual information available in a digital format, effective retrieval is difficult without good indexing and summarization of the document content. Document categorization is one solution to this problem and it is a very common automatic mechanism for information extraction. This is a task of automatically assigning predefined categories to free text documents [2].

A growing number of statistical classification method and Machine Learning (ML) techniques have been applied to document categorization in recent years, including Multivariate Regression Models, Nearest Neighbor Classification, Bayes Probabilistic Approaches, Symbolic Rule Learning, and Inductive Learning Algorithm.

Fuzzy clustering has also been used for document categorization [3], [4]. Document clustering has also been widely applied in the field of information retrieval for improving search and retrieval efficiency.

The dimensionality problem, as questioned by Yang and Pedersen in [2] more than ten years ago, has been one of the topics frequently discussed in the ML community in the last years. This important issue continues to attract the attention of researchers nowadays.

One of the major characteristics or difficulties of the document categorization is the high dimensionality of the feature space. The native feature space consists of the unique terms that occur in the documents, which can be thousands of terms for even a moderate-sized text collection. Thus, documents are represented as sets of terms, in which the terms constitute the conceptual units for describing user's information needs.

The term dimensionality is assigned to the number of features in a pattern representation, i.e. the dimension of the feature space. The two main reasons for the dimensionality to be as small as possible are: the cost of the measurement and the accuracy of the classifier. Therefore, a reduced number of features can prevent the so-called Curse of Dimensionality [5], which refers to the exponential growth of the hyper volume as a function of dimensionality.

Furthermore, automatic information extraction and access from large data sets, especially data sets obtained from text documents, faces an important issue that is the scaling up problem. This problem occurs during the evaluation of the document classification algorithm. Besides producing excessive storage requirements, the scaling up increases time complexity and affects to generalization accuracy, introducing noise and overfitting [6].

In this work, we consider the issues of dimensionality and scalability of a document classifier in terms of information accessing. The classification task is one of the steps to automatic recover the information that the user accesses by a query in an Information Retrieval System (IRS).

We are addressing the dimensionality reduction of text collections by means of fuzzy groups and classification of documents through the construction of a fuzzy rule base obtained from these groups.

This paper is organized as follow: in Section 2, we discuss the dimensionality problem in text information retrieval. Next, in Section 3 we introduce the experimental methodology for classification that is considered in this research followed by some results in Section 4. Finally, in Section 5, we conclude and point future directions of this research.

II. The Dimensionality Problem in Text Information Retrieval

According to Herrera-Viedma in [7], the Information Retrieval (IR) involves the development of computer systems for the storage and retrieval of textual information (documents). The main activity of an IRS is to gather pertinent documents that better satisfy the user information requirements (query).

According to Bordogna et. al. in [8], by formulating a request for information, an user tries to express a set of concepts, that considers essential, to retrieve that information - this process is subjective and imprecise in nature. When using an IRS, an user engages with a collection of stored information through an automatic intermediary. The automatic intermediary analysis the user's request and retrieves the information that it judges to be satisfying for the user's needs. This request must be expressed in a query language, which the automatic intermediaries are able to interpret. The ability of systems to interpret a request begins in the representation of the documents. Thus, it is essential that the method of document classification aim the comprehensibility of the documents during the information retrieval process.

During a request, it is also important to treat the term relevance according to the user's preference. In this sense, Pasi and Bordogna in [8]–[10] were the first to propose the treatment of the term relevance in a query as a fuzzy linguistic variable and, according to them, the Fuzzy Set Theory (FST) proposed by Zadeh in [11] provides a simple and suitable means to deal with qualitative and imprecise criteria.

FST enables the interpretation of a certain knowledge expressed in a linguistic format through a mathematical representation. In general, the use of fuzzy sets can consider the imprecise knowledge inherent to the real world. The fuzzy sets also allow the representation of vague concepts expressed through linguistic terms such as high temperature, low cost or cold weather [12].

It is possible to develop more powerful mechanisms to represent knowledge by exploring the use of Fuzzy Classification Systems (FCS) and managing a level of details of knowledge. These details should be sufficient to deal with vagueness and uncertainty of real knowledge, because they include the management of uncertainty in the final model.

FCS are Fuzzy Rule-Based Systems (FRBS) designed with the specific goal to perform the task of classification. A fuzzy rule is an if-then rule that describes a chunk of knowledge about a domain using propositions of Fuzzy Logic [12] by establishing relations between input variables and output variables.

In FRBS, the high dimensionality with large number of features can be tackled from a double perspective [13]:

1. Through the compactness and reduction of the rule set, minimizing the number of included fuzzy rules. Unnecessary rules can be eliminated with the aim of having a more cooperative rule set in order to obtain an FRBS with better performance.
2. Through a feature selection process that reduces the operation level among the rules. The operation level depends on the number of features used by the FRBS.

Some aspects of imprecision of the textual information have been studied in the literature using the theory of fuzzy sets combined with Clustering. These studies cover the development of clustering techniques for the organization of documents in a Text Mining (TM) process, and IR and focus mainly on improving the ways of indexing or querying of documents.

Clustering techniques have been applied since long time to organize documents belonging to the same cluster in adjacent positions in the storage media, thus minimizing the number of accesses needed to retrieve documents about a topic. According to Rodrigues and Sacks in [14], topics that characterize a given domain of knowledge are sometimes associated with each other. Furthermore, these topics may also be related to the topics of different areas and the documents may contain relevant information to differentiate fields to some degree.

Some approaches using fuzzy clustering algorithms for TM can be found in [15]–[19]. Through these algorithms, documents are assigned to multiple groups simultaneously and relationships among the domains can be found.

In addition to the effort made by researchers to solve the problems of TM using fuzzy clustering, researchers in the field of IR have been developing models for the representation of large collections of text documents. According to Crestani and Pasi in [20], the effectiveness of information retrieval systems is related to the ability of the systems to deal with uncertainty and imprecision of the recovery process. However, commercially available IRS ignore this type of information, by simplifying the representation of the document contents and the interaction of the user with the system.

In recent years, a great deal of research in IR has aimed at modeling the vagueness and uncertainty, which invariably characterize the management of information. The most common approaches are those based on methods of Natural Language Processing (NLP) analysis [20], Probabilistic Methods (PM) [21], and Soft Information Retrieval (SIR) [22].

The main limitation of methods based on NLP is the deepness level of the language analysis, and their consequent range of applicability. A satisfying interpretation of the meaning of the document needs a very large number of decision rules even in narrow application domains. On the other hand, approaches based on PM are more general: their objective is to define retrieval models, which deal with imprecision, and uncertainty independently on the application domain. The aim of probabilistic IR is to develop ad hoc models able to deal with the uncertainty of the retrieval process. Furthermore, the set of approaches based on IRS has received increasing interest from the use of techniques to deal with vagueness and uncertainty. The application of soft computing to IR, especially FST, is particularly useful to model mechanisms which learn the user notion about relevance of documents [21].

Therefore, the FST has been successfully employed to model relevance judgment formulations based on the extension of document representation [12]. In addition to this, according to Kolland and Srinivasan in [23], it has proved that FST presents better results than probabilistic models.

The work of Bordogna and Pasi in [24] is the most relevant to our approach. In this paper, the authors show a good motivation for linguistic representation of documents, suggesting the use of rules for information retrieval, qualifying linguistically both the terms that represent the document and the retrieved documents. But in their approach they do not consider the dimensionality and interpretability problem of the representation of documents. The interpretability problem in fuzzy classification systems has been improved by a lot of methods [25]–[32]. On the other side, the dimensionality problem is focused in this paper.

A FRBS exponentially increases the size of the input and output space caused by the use of linguistic variables when problems with high dimensionality are addressed. For example, problems in which the knowledge obtained from a collection of textual documents is used. A large number of variables do not guarantee a good interpretability. Moreover, with an excessive number of input variables, the linguistic rules also lose part of its description ability, and the user's understanding about the condition to activate the rule becomes more difficult.

Approaches focusing on the same topic presented in this paper aim to bring a group of documents closer to the real relevance of a document, which is related to a particular topic of a specific collection [4], [15], [16]. The relevance is measured by the expectation of the user when organizing and classifying a particular document.

Given this premise, we developed a mechanism for the generation of classification fuzzy rules from a given document collection so that we can classify the documents using a smaller search space. This mechanism is explained in the next section.

III. Fuzzy Rule-Based Documents Classification

Vagueness and uncertainty are present in all textual information, as different writers or readers deal with the text from different perspectives and representation of the content (when organizing documents) or expectations (for a query). Moreover, generally, the decision or classification of the relevance of these texts is associated with a given condition imposed by the user or even the words and terms that were used throughout the text. Fuzzy logic can lead to techniques that deal with the vagueness and uncertainty typical of real situations, besides being widely used and well founded.

With the use of fuzzy clustering, specifically, the documents can belong to more than one domain topic, represented by one group, with varying degrees of relevance favoring the organization of textual information. In general, in the pattern extraction step by means of clustering algorithms on the text mining process, topics related to the documents are inferred from lists of words from the text that are the most discriminating of each cluster.

Using fuzzy rules from fuzzy clustering, the relevance of documents with relation to groups can be represented by means of linguistic terms, which resembles in a more appropriate way the indication of importance given by human beings. For example, a document can belong "a lot", or "a little" to a particular group/topic, or a topic can be "very important", or "less important" for the user's query. Moreover, the importance of documents in groups through linguistic terms allows the generation of fuzzy rules that can be used in the recovery of textual information, through the reasoning mechanisms that use fuzzy rules with facts, to make inferences.

With this motivation, we propose a dimensionality reduction using fuzzy clustering of documents and then, the generation of classification fuzzy rules.

A. Dimensionality reduction

After filtering the base of documents, the preprocessing of texts should seek to structure the documents in order to make them able to be analyzed by the algorithms of pattern extraction. The most common textual data preprocessing is a representation of a vector space in the form of an attribute-value matrix, so that each line corresponds to one document in the collection and each column corresponds to one term (attribute) in this entire collection of documents. Thus, each cell of the matrix is associated with a measure such as the binary measure, indicating the presence or absence of a term in a document, the frequency of a term in a document, or the weighted frequency of a term in a document according to its distribution throughout the collection.

Generally, the terms present in the attribute-value matrix are first examined and prepared. In an initial effort, we seek to disregard terms that do not represent useful knowledge through the elimination of stopwords, which are not relevant words in the analysis of texts and usually consist of prepositions, pronouns, articles, and interjections, among others.

The attribute-value matrix of the document collection is inherently high dimensional and sparse, which sometimes can make the process of analysis computationally very expensive or even impossible. This negatively affects the outcome of some knowledge extraction algorithms. Thus, it is very important for the process of analysis to select the most relevant terms from the document collection, making the set of terms more concise but not less representative in relation to the original set.

Furthermore, according to Sebastiani in [33], there is a risk in the removal of terms, since it is possible to remove potentially useful information about the meaning of the documents. Thus, in order to obtain optimal (cost) effectiveness, the reduction process must be performed with care.

There are two ways of dimensionality reduction in terms of the nature of the resulting terms [33]: (1) Term selection, in which the set of term T' is a subset of T, and (2) Term extraction, in which the terms in T' are not of the same type of the terms in T, but are obtained by combinations or transformations of the original ones. For example, if the terms in T are words, the terms in T' may not be words at all.

In this work, we manage the dimensionality reduction with a term selection by frequency and propose a new term extraction method.

An attribute-value matrix is considered as a document-term matrix $A_{\{m\times n\}}$, in which each cell a_{dt} is represented by the frequency f_{dt}, i.e., the frequency (value) of the term (attribute) t in a document d.

In general, the text classification consists of a set of classes $C = \{c_0, c_1, \ldots, c_k\}$, $k \in$ N, a set of documents D where each document is represented by a vector of frequencies $\boldsymbol{F}$, and a function r: $D \rightarrow \{0, 1, \ldots, k\}$ called classification rule. D is the set of all documents so that each document $d \in D$ belongs to exactly one class in C. Thus, it is desired to find a set of S classification rules that returns the correct class label for each document in D.

If the rules were generated from an attribute-value matrix without reduction, the interpretability would be incomprehensible. For example: suppose there is a collection of 100 documents and, from the preprocessing, 1000 terms are obtained that represent these documents. Considering that each term is represented by a variable in the rule, a rule with 1000 variables in the antecedent is not easy to interpret by human beings. Furthermore, an information retrieval process could spend a lot of time to decide on the best document to be recovered for the user. Therefore, with a big number of antecedents in the rule, the search space increases exponentially.

In order to reduce the number of antecedents in the rules without the loss of information by the discard of terms, we proposed a method for term extraction transforming the document-term matrix $A_{\{m\times n\}}$, which represents documents by the frequency of terms, in a matrix with a minor dimensionality, as a document-group matrix $A'_{\{m'\times n'\}}$.

For the purpose of the dimensionality reduction of the document-term matrix and better process of information retrieval, we modified the Fuzzy C-Means algorithm to cluster text documents into groups so that all documents belong to all groups with different membership degrees. The change in the clustering algorithm was obtained replacing the original Euclidean distance measure used in the Fuzzy C-Means by the Cosine similarity to manage the text documents because of the large dimensionality of the attribute-value matrix.

The membership degrees represent the relevance of a document in each group and compose the cell of the obtained new matrix. Therefore, we work with an attribute-value matrix A', with $A'_{\{m'\times n'\}} < A_{\{m\times n\}}$, i.e. $m' = m$ e $n' < n$, in which each cell a_{dg} is represented by the membership degree $A_i(a_{dg})$, i.e., the membership degree (value) of the document (attribute) d in a group g.

B. Classification fuzzy rules

Since the groups have been found, we can generate the fuzzy rules by any fuzzy rule generation method. In this work we used the well-known Wang&Mendell [34] method because of its simplicity and good results discussed on the literature. The generated rules can then be used to classify the documents.

In our proposal, the rules will assume the format:

IF G_1 is *Low* **AND** G_2 is *Mid* **AND** G_3 is *High* **THEN**
C is *Artificial Intelligence*

where G_1, G_2 ... G_n represent groups formed by clustering the documents and C is the class of the documents.

The inference method used in this paper works as follows. Let $d_p = (g_{p1}, g_{p2}, \ldots, g_{pn})$ be a document to be classified, in which $(g_{p1}, g_{p2}, \ldots, g_{pn})$ represents the membership degree of the document d_p in each group, and $\{R_1, R_2, \ldots, R_S\}$ the set of S rules of the classification system, each one with n antecedents. Let $A_i(g_{pi})$, $i = 1, \ldots, n$, be the membership degree of the attribute g_{pi} in the i-th fuzzy group of the rule R_k. Based on the Classical Fuzzy Reasoning (CFR) method, also known as the "winner rule" method [35], the process of inference used to classify the document d_p is:

1. Calculate the compatibility degree between the input document d_p and each rule R_k, $k = 1,.., S$:
   $\mathrm{C_{ompat}}(R_k,d_p)=t(A_1(g_{p1}),A_2(g_{p2}),\ldots, A_n(g_{pn}))$, in which t denote one t-norm;
2. Find the rule R_{kmax} with the highest compatibility degree with the document;
3. Define as output of the inference process the class C_k of the rule found in the previous step.

IV. Experimental Results and Analysis

In previous experiments presented in [36], we analyzed the influence of the preprocessing on the generation of fuzzy rules by clustering, because at the end of the preprocessing step the attribute-value matrix obtained is inherently high dimensional and sparse. Such characteristics sometimes can make the process of analysis computationally very expensive or even impossible, and negatively affect the outcome of some algorithms for knowledge extraction.

In these experiments, five data sets obtained from the proceedings of the ACM digital library (http://portal.acm.org/) were used. Each data set has 5 classes with about 90 documents (instances) per class. These data sets were preprocessed using the Pretext tool [37] in order to be converted to an attribute-value matrix that contains the frequencies of each term/word (attributes) in a document.

Also using the Pretext tool, it was possible to reduce the number of terms through a feature selection by the frequency. The feature selection step was done according to the five different conditions (tests) shown in Table 1, which were defined varying the minimum and maximum frequencies of selected terms. These particular frequency values were set with the objective of designing testing conditions varying both the minimum frequencies and the number of attributes, considering that our aim was to analyze how much it is possible to carry out a feature selection without loss of information, measured by the rate of correct classification. For instance, in the test 1, all the terms that occur in the documents with frequencies between 50 and 500 were selected.

	Frequencies	
Tests	Minimum	Maximum
1	50	500
2	100	300
3	50	100
4	500	1000
5	50	1000

Table 1. Feature selection by frequency.

With these tests presented in [36], we checked whether the variation in the quantity of attributes (Table 2) derived from the attribute selection parameters (Table 1) interposes in the performance of the proposed method.

Domain	Instances	Test1	Test2	Test3	Test4	Test5
Exp2	399	3132	1357	1436	1713	3398
Exp2	410	2722	1166	1299	1442	2945
Exp3	424	3073	1371	1356	1741	3326
Exp4	394	3072	1313	1430	1653	3352
Exp5	471	3471	1522	1577	1916	3807

Table 2. Number of features in each test.

We carry out a comparison among the obtained results by the proposed method in each test for each domain, represented by the chart in Figure 1. In this chart, each line represents a domain and the vertical axis represents the correct classification rate obtained in each test. In order to place multiple lines on the same chart, but separated from each other to be visualize all at once, a Stacked Line Chart was used, which displays the trend of the contribution of each value over categories.

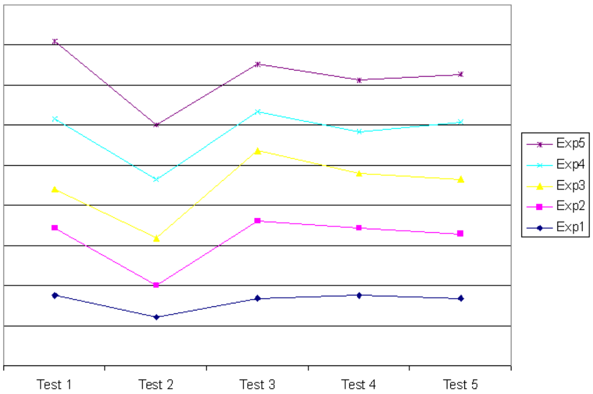

Figure 1. The results obtained by the proposed method in each test.

As observed in Figure 1, the preprocessing that most interposes in the results by lowering the quality of the classification, was the feature selection in Test 2. In this test the number of attributes was much reduced compared to the others tests. The results obtained from the experiments suggest that the drastic reduction in the number of attributes led to a loss of information. Thus, we decided to continue the experiments with the selected domains considering the feature selection by frequency obtained in Test 3, which presented an intermediate value among the number of selected features in each test. Furthermore, in Test 3 there is not a significant loss in the correct classification rate obtained by the proposed method compared with the other methods.

We improved the analysis started in [36], comparing the proposed method with the well-known document classification method: Support Vector Machine (SVM). Other classification methods were also compared: OneR, K-Nearest Neighbors classifier (K-NN), C4.5 decision tree algorithm and Naive Bayes.

The SVM constructs a hyperplane or set of hyperplanes in a high or infinite dimensional space, measuring the complexity of hypotheses based on the margin with which they separate the data, not the number of features. Thus, the biggest advantage of SVM is its ability to learn independent of the dimensionality of the feature space. However, according to Shanahan and Roma in [38], the SVM, when applied to text classification, provides excellent precision, but poor recall.

OneR is a machine learning algorithm which produces very simple rules based on a single attribute. On the other hand, the C4.5 algorithm [39] is a predictive machine-learning model that decides the target value of a new sample based on various attribute values of the available data.

The results obtained by Holte in [40] indicate that simple modifications to oneR might produce a system competitive with C4.5. Furthermore, the comparison among oneR and other algorithms aims the evaluation of the tradeoff between simplicity and accuracy.

The K-Nearest Neighbor (K-NN) is an Instance-Based Learning (IBL) method. IBL approaches can construct a different approximation to the target function for each distinct query instance that must be classified. In fact, the K-NN constructs a local approximation to the target function that applies in the neighborhood of the new query instance, and never constructs an approximation designed to perform well over the entire instance space. This has significant advantages when the target function is very complex. However, the disadvantage to K-NN is that it typically considers all attributes of the instances when attempting to retrieve similar training examples from memory [41].

Experimental results obtained by Joachims in [42] show that SVMs consistently achieve good performance in text categorization tasks, outperforming the other compared methods. However, Gabrilovich and Markovitch in [43] demonstrate that in such datasets C4.5 significantly outperforms SVM and KNN, although the latter are usually considered substantially superior to text classifiers. According to the authors, when no feature selection is performed, C4.5 constructs small decision trees that capture the concept much better than either SVM or KNN. Furthermore, even when feature selection is optimized for each classifier, C4.5 formulates a powerful classification model, significantly superior to that of KNN and only marginally less capable than that of SVM.

The Naive Bayes classifier is based on the Bayes rule of conditional probability. It uses all the attributes contained in the data, and analyses them individually. According to Schneider in [44], this method is often used in text classification applications and experiments because of its simplicity and effectiveness. However, its performance is often degraded because it does not model the text very well.

A lot of experiments have been carried out comparing all these methods in order to analyze their performance for text categorization. With this motivation, and to show that our method is comparable with the state-of-art, we carried out some experiments comparing the correct classification rates obtained from our method and the well-known machine learning methods for classification.

The methods used for comparisons in the experiments were applied in the documents frequency matrix, without the previous clustering step. Furthermore, the classification was done with the default parameters of each method on WEKA tool [45].

Attempting to achieve an estimate error close to the true error, the 10-fold cross validation method was used in all experiments and we present the obtained results in Table 3.

To test whether there is a significant difference among the methods, the Friedman with Nemenyi post-hoc tests were used with the null-hypothesis that the performance of the six methods, assessed in terms of correct classification rates, are comparable.

The critical value of the Chi-square statistics with 5 degrees of freedom is 11.07. Thus, according to the Freidman test using the Chi-Square statistics, the null-hypothesis that all algorithms behave similar was rejected with a 95% of confidence level. Furthermore, according to the Nemenyi statistics, the critical value for comparing the mean ranking of two different algorithms at 95% of confidence level is 3.37. Mean-rankings differences greater than this value are significant.

In Figure 2, the overall comparison among feature ranking methods is presented. Methods are ordered by performance from left to right. A thick line joining two or more methods indicates that there are no significantly difference methods. Therefore, it is possible to notice that the Naïve Bayes, SVM, C4.5 and K-NN methods are not significantly different of the proposed method.

	Proposed Method	**K-NN**	**C4.5**	**Naïve Bayes**	**OneR**	**SVM**
Exp1	0.840(1.0)	0.371(5.0)	0.719(4.0)	0.802(2.0)	0.318(6.0)	0.789(3.0)
Exp2	0.960(1.0)	0.505(5.0)	0.844(3.0)	0.827(4.0)	0.410(6.0)	0.878(2.0)
Exp3	0.880(1.0)	0.238(6.0)	0.696(4.0)	0.792(2.0)	0.307(5.0)	0.762(3.0)
Exp4	0.480(5.0)	0.485(4.0)	0.838(3.0)	0.914(1.0)	0.396(6.0)	0.858(2.0)
Exp5	0.600(4.0)	0.323(5.0)	0.730(3.0)	0.837(1.0)	0.291(6.0)	0.824(2.0)
Average Rank	2.400	5.000	3.400	2.000	5.800	2.400

Table 3. Correct classification rate.

In Figure 2, the overall comparison among feature ranking methods is presented. Methods are ordered by performance from left to right. A thick line joining two or more methods indicates that there are no significantly difference methods. Therefore, it is possible to notice that the Naïve Bayes, SVM, C4.5 and K-NN methods are not significantly different of the proposed method.

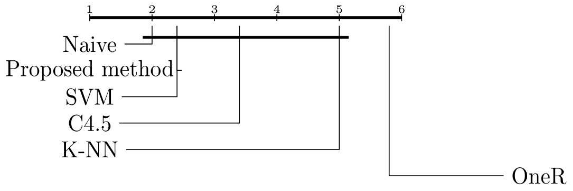

Figure 2. Comparison among feature ranking methods.

The results showed that the proposed method is comparable with the well-known machine learning methods for classification and even comparable with the SVM method for document classification.

The comparative analysis was done through the accuracy of each method. However, an important aspect to highlight in this contribution is that the dimensionality reduction is very important for the use of fuzzy rules in IR problems. A feature selection process reduces the operation level among the rules, because the operation level depends on the number of features used by the Fuzzy Rule-Based Systems.

The use of fuzzy rules by IRS is highly beneficial when the relevance values cannot be expressed by means of numerical values. The linguistic approach of IRS gives a more natural way to the user requests his/her needs.

V. Conclusions and Perspectives

A wide number of algorithms and proposals of the state-of-the-art of machine learning have been discussed in the field of document categorization. In this article, we presented a comparison analysis between some of the main machine learning classification algorithms for document classification and our proposed method.

We developed a mechanism to reduce the high dimensionality of the attribute-value matrix obtained from the documents. The results obtained showed that our method is comparable to some of the most effective classification method in machine learning and statistics fields. In addition, the proposed classifier scaled the document classification up, reducing the operation level among fuzzy rules.

Furthermore, this paper has considered the dimensionality reduction to tackle an important issue in the development of Information Retrieval Systems: to recover a relevant document for a given user's query using a small search space. Thus, the experiments showed that this is a promising approach to deal with the problem of dimensionality when fuzzy rules are used to classify documents.

Moreover, the Wang&Mendell method is a grid-based method and one of its problems is the large number of rules produced. However, for the proposed method the number of rules was very good, because the number of rules were about 5% of the number of documents.

Investigations will continue in the future including the experiments with different domains, preprocessing of the documents and analysis of the reduction in the number of rules. Currently, the authors have been working on an extension of the method to carry out information retrieval after the document organization using rules, in order to make inference considering that the same document can activate different rules but with different degrees.

Acknowledgment

The authors would like to thank the CAPES (Coordination for the Improvement of Higher Level – or Education – Personnel) and CNPq (National Counsel of Technological and Scientific Development) Brazilian research agency for their financial support.

References

[1] E. Herrera-Viedma, HA.G. López-HerreraH. "A Review on Information Accessing Systems Based on Fuzzy Linguistic Modeling". *International Journal of Computational Intelligence Systems*, v. 3(4), 2010, pp. 420-437.

[2] Y. Yang and J. O. Pedersen. "A Comparative Study on Feature Selection in Text Categorization". In *Proceedings of the 14th International Conference on Machine Learning*, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1997, pp. 412-420.

[3] E. M. Rodrigues and L. Sacks. "Learning topic hierarchies from text documents using a scalable hierarchical fuzzy clustering method". In *Proceedings of the International Conference on Recent Advances in Soft Computing*, 2005, pp. 269-274.

[4] R. Saracoglu, K. TuTuncu and N. Allahverdi. "A new approach on search for similar documents with multiple categories using fuzzy clustering", *Expert Systems with Applications*, v. 34, 2008, pp. 2545-2554.

[5] R. Bellman. *Adaptive Control Processes: A Guided Tour*, Princeton University Press, 1961.

[6] J.R. Cano, F. Herrera and M. Lozano. "Stratification for scaling up evolutionary prototype selection", *Pattern Recognition Letters*, v. 26 (7), 2005, pp. 953-963.

[7] E. Herrera-Viedma. "Modeling the retrieval process for an information retrieval system using an ordinal fuzzy linguistic approach", *Journal of the American Society for Information Science and Technology*, v. 52(6), 2001, pp. 460-475.

[8] G. Bordogna, P. Carrara, and G. Pasi. "Extending Boolean information retrieval: a fuzzy model based on linguistic variables". In *Proceedings of the First IEEE International Conference on Fuzzy Systems*, 1992, pp. 769-776.

[9] G.Bordogna, P.Carrara and G.Pasi. "Query term weights as constraints in fuzzy information retrieval", *Information Processing and Management*, v. 27(1), 1991, pp.15-26.

[10] G.Bordogna and G.Pasi. "A fuzzy linguistic approach generalizing Boolean information retrieval: a model and its evaluation", *Journal of the American Society for Information Science*, v. 44(2), 1993, pp. 70-82.

[11] L. A. Zadeh. "Fuzzy sets", *Information and Control*, v. 8(3), 1965, pp. 338-353.

[12] G. J Klir and B. Yuan. *Fuzzy Sets and Fuzzy Logic: theory and applications*, 1 ed., Prentice-Hall, 1995.

[13] O. Cordon, F. Herrera, M. J. del Jesus and P. Villar. "A multiobjective genetic algorithm for feature selection and granularity learning in fuzzy-rule based classification systems". In *Proceedings of the Joint 9th IFSA World Congress and 20th NAFIPS International Conference*, v.3, 2001, pp. 1253-1258.

[14] D. Angluin and P. Laird. "Learning from noisy examples", *Machine Learning*, v. 2(4), 1987, pp. 343-370.

[15] V. Torra. "Fuzzy c-means for fuzzy hierarchical clustering". In *Proceedings of the IEEE International Conference on Fuzzy Systems*, 2005, pp. 646-651.

[16] K.M. Lee. "Mining generalized fuzzy quantitative association rules with fuzzy generalization hierarchies". In *Proceedings of the Joint 9th IFSA World Congress and 20th NAFIPS International Conference*, v. 5, 2001, pp. 2977-2982.

[17] Y. -J. Horng, S. -M. Chen, Y.-C. Chang and C. -H. Lee. "A new method for fuzzy information retrieval based on fuzzy hierarchical clustering and fuzzy inference techniques", *IEEE Transactions on Fuzzy Systems*, v. 13(2), 2005, pp. 216-228.

[18] G. Bordogna, M. Pagani and G. Pasi. *Soft Computing or Information Retrieval on the Web*, Springer Verlag, 2006.

[19] G.Bordogna and G.Pasi. "Hierarchical-hyperspherical divisive fuzzy c-means (h2d-fcm) clustering for information retrieval". In *Proceedings of the IEEE International Conference on Web Intelligence and Intelligent Agent Technology*, 2009, pp. 614-621.

[20] A.F. Smeaton. "Progress in the application of natural language processing to information retrieval tasks", *Computer Journal*, v. 35, 1992, 268-278.

[21] F. Crestani, M. Lalmas, C.J.V. Rijsbergen and I. Campbell. "Is this document relevant? ...probably: A survey of probabilistic models in information retrieval", *Journal ACM Computing Surveys*, v. 30(4), 2001, pp. 528-552.

[22] F. Crestani and G. Pasi. "Soft information retrieval: Applications of fuzzy set theory and neural networks", *Neuro-fuzzy Techniques for Intelligent Information Systems*, N.Kasabov and R. Kozma, Eds. Physica-Verlag, Springer- Verlag Group, 1999, pp. 287-313.

[23] M. Koll and P. Srinivasan. "Fuzzy versus probabilistic models for user judgments", *Journal of the American Society for Information Science*, v. 4(4), 1990, pp. 264-271.

[24] G. Bordogna and G. Pasi. "Fuzzy rule based information retrieval". In *Proceedings of the 18th International Conference of the North American Fuzzy Information Processing Society*, NAFIPS, 1999, pp. 585-589.

[25] S. Guillaume. "Designing fuzzy inference systems from data: An interpretability-oriented review", *IEEE Transactions Fuzzy Systems*, v. 9(3), 2001, pp. 426-443.

[26] J. Abonyi, J. A. Roubos and F. Szeifert. "Data-driven generation of compact, accurate, and linguistically sound fuzzy classifiers based on a decision-tree initialization", *International journal of approximate reasoning*, v.32, 2003, pp. 1-21.

[27] H. Wang, S. Kwong and Y. Jin. "A multi-objective hierarchical genetic algorithm for interpretable rule-based knowledge extraction", *Fuzzy Sets and Systems*, v. 149(1), 2005, pp. 149-186.

[28] T.M. Nogueira and H. A. Camargo. "Fuzzy-CCM: a context-sensitive approach to fuzzy modeling", *International Journal of Hybrid Intelligent Systems*, v. 7, 2010, pp. 33-43.

[29] T. M. Nogueira and H. A. Camargo. "Conditional clustering and the generation of fuzzy models". In

Proceedings of the II Workshop on Computational Intelligence, 19th Brazilian Symposium on Artificial Intelligence, 2008, pp. 63-68.

[30] T. M. Nogueira and H. A. Camargo. “Context-sensitive clustering in the design of fuzzy models”. In *Proceedings of the International Conference on Hybrid Intelligent Systems*, 2008, pp. 240-245.

[31] T. M. Nogueira and H. A. Camargo. “Fuzzy rule base generation through conditional clustering”. In *Proceedings of the IV Workshop on MSc Dissertation and PhD Thesis in Artificial Intelligence, 19th Brazilian Symposium on Artificial Intelligence (SBIA08)*, 2008, pp. 1–10.

[32] J. Casillas, O. Cordón, F. Herrera and L. Magdalena. “Interpretability improvements to find the balance interpretability-accuracy in fuzzy modeling: an overview”, *Chapter of Interpretability Issues in Fuzzy Modeling*, Springer, 2003, pp. 3-22.

[33] F. Sebastiani. “Machine learning in automated text categorization”, *ACM Computing Surveys*, v. 34(1), 2002, pp. 1-47.

[34] L. Wang and J. Mendel. “Generating fuzzy rules by learning from examples”, *IEEE Transaction on Fuzzy Systems, Man and Cybernetics*, v. 22, 1992, pp. 414-1427.

[35] Z. Chi, H. Yan and T. Pham. *Fuzzy Algorithms with Applications to Image Processing and Pattern Recognition*, World Scientific, 1996.

[36] T. M. Nogueira, H. A. Camargo and S. O. Rezende. “On the use of fuzzy rules to text document classification”. In *Proceedings of the 10th International Conference on Hybrid Intelligent Systems*, 2010, pp. 19-24.

[37] M. V. B. Soares, R. C. Prati and M. C. Monard. “PreTexT II: Description of restructuring tool preprocessing of texts”, *Technical report* 333, ICMC-USP, Brazil, 2008 (in portuguese).

[38] J. Shanahan and N. Roma. “Improving SVM Text Classification Performance through Threshold Adjustment”, *Machine Learning, Lecture Notes in Computer Science*, 2003, v. 2837, pp. 361-372.

[39] J. R. Quinlan, *C4.5: Programs for machine learning*, Morgan Kaufmann, 1993.

[40] R. C. Holte. “Very Simple Classification Rules Perform Well on Most Commonly Used Datasets”, *Machine Learning. Lecture Notes in Computer Science*, 1993, v. 11(1), pp. 63-90.

[41] T. Mitchell. *Machine Learning*, McGraw-Hill Education, 1 edition, 1997.

[42] T. Joachims. “Text categorization with Support Vector Machines: Learning with many relevant features”, *Machine Learning. Lecture Notes in Computer Science*, 1998, v. 1398, pp. 137-142.

[43] E. Gabrilovich and S. Markovitch. “Text Categorization with Many Redundant Features: Using Aggressive Feature Selection to Make SVMs Competitive with C4.5”. In *Proceedings of the 21th International Conference on Machine Learning*, 2004, pp. 321-328.

[44] K. Schneider. “Techniques for Improving the Performance of Naïve Bayes for Text Classification”, *Computational Linguistics and Intelligent Text Processing, Lecture Notes in Computer Science*, 2005, v. 3406, pp. 682-693.

[45] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten. “The WEKA data mining software: An update”, *SIGKDD Explorations*, 2009, v. 11(1).

Author Biographies

Tatiane Marques Nogueira was born in Feira de Santana, Bahia, Brazil, in 1982. She is graduated in Computer Science (2006) from the State University of Southwest Bahia (UESB), master in Artificial Intelligence (2008) from the Federal University of São Carlos (UFSCar) and is currently a PhD student in Computational Intelligence at the University of São Paulo, ICMC-USP São Carlos. She has experience in Computer Science with emphasis on Logic and Semantics of Programs, acting on the following topics: Fuzzy Logic, Clustering and Text Mining.

Heloisa de Arruda Camargo was born in São Carlos-SP, Brasil, in 1956. She is graduated in Computer Science and master in Computer Science (1984) from ICMC-USP São Carlos, and PhD in Electrical Engineering (1993) from FEEC-UNICAMP, Campinas-SP. She is an associate professor at the Department of Computer Science, Federal University of São Carlos (UFSCar), where she has been working since 1980. In the period from 2001 to 2002, she did postdoctoral work at the University of Alberta, Edmonton, AB, Canada. Her main research lines are: Genetic Fuzzy Systems, Semi-supervised Learning, Reasoning and Approximate Methods for Hybrid Fuzzy Modeling.

Solange Oliveira Rezende is graduated in Mathematics from the Federal University of Uberlandia (1986), Master in Computer Science and Computational Mathematics from the University of São Paulo (1990) and Ph.D. in Mechanical Engineering from the University of São Paulo, São Carlos (1993). She is currently an associate professor at the University of São Paulo. She has experience in Computer Science with emphasis in Computer Methods and Techniques, and Artificial Intelligence, working mainly on issues related to data mining and text.