Document Classification for Large Datasets Based On Hesitant Fuzzy Linguistic Term Set

Swatantra Kumar Sahu 1, Bharat Mishra2 and R. S. Thakur3
1 M.G.C.G.V.V., Satna, Madhya Pradesh, India.
2 M.G.C.G.V.V., Satna, Madhya Pradesh, India.
3 MANIT, Bhopal, Madhya Pradesh, India.

Abstract

This paper presents Hesitant Fuzzy information about large data sets. Hesitant Fuzzy Linguistic Term Set (HFLTS) is based on the fuzzy linguistic approach that will serve as basis to Increase the flexibility of elicitation of linguistic Information. Experimental results evaluated using the Analytical Tool MATLAB 7.14. The classification results show the proposed approach performs well.

Keywords: Hesitant Fuzzy Set, Classification, Large Data sets, Linguistic Term Set.

1. INTRODUCTION

Hesitant Fuzzy Information collection is refer Fuzzy logic, Fuzzy sets theory, Intuitionistic fuzzy sets, Fuzzy multi sets, fuzzy linguistic approach, uncertainty and loading of data etc. In this Paper Hesitant Fuzzy Linguistic Term Set (HFLTS) is used to classify the document datasets.

![Fig 1: Hesitant Fuzzy Information collection](image-url)
In the area of document classification various approach proposed by researcher some of listed blow Swatantra Kumar Sahu, et.al.“Hesitant Fuzzy Linguistic Term Set Based Document Classification”[48], S.A. Orlovsky, “Decision-making with a fuzzy preference relation[25], Swatantra Kumar Sahu, et.al “Numerical Result Analysis of Document Classification for Large Data Sets” [49],H. Becker, “Computing with words and machine learning in medical diagnosis[2],Y.Dong,et.al.“Computing the numerical scale of the linguistic term set for the 2-tuple fuzzy linguistic representation model [5].


H. Ishibuchi, et.al. , Classification and Modelling With Linguistic Information Granules: Advanced Approaches to Linguistic Data Mining [13],D. F. Li, “Multi attribute group decision making method using extended linguistic variables[17].


relations[28].


D.Wu et.al., “Computing with words for hierarchical decision making applied to evaluating a weapon system [36], A. Sengupta et.al., “On comparing interval numbers [29].


S. M. Zhou, R et.al., “On aggregating uncertain information by type-2 OWA operators for soft decision-making[43], S. Parsons, “Current approaches to handling imperfect information in data and knowledge bases[26]. Hesitant Distance Similarity Measures for Document Clustering[45].


This paper is organized as follows. Section-1 described the introduction and review of literatures. In Section-2, the Hesitant Fuzzy Information is described. In Section-3, Methodology of document Classification is described. In Section-4, Experimental results are described. Finally, we concluded and proposed some future directions in Conclusion Section, i.e. Section 5.

2 HESITANT FUZZY LINGUISTIC TERM SET

Uncertainty problem is occurs during calculation of document classification results, for handling this problem the best and optimum solution is Hesitant Fuzzy Set. Hesitant Fuzzy Set gives new computational solution with numerical capability. Hesitant Fuzzy used Linguistic Term Set it knows Hesitant Fuzzy Linguistic Term Set (HFLTS). Linguistic Term Set just like Context Free Grammar (CFG) [44].

Definition 1: Let S be a linguistic term set, $S=\{S_0,\ldots,S_g\}$, an HFLTS, $H_s$, is an ordered finite subset of the consecutive linguistic terms of S. and define the empty HFLTS and the full HFLTS for a linguistic variable $\lambda$ as follows.
1) Empty HFLTS: \( H_0(\lambda) = \{ \} \)

2) Full HFLTS: \( H_1(\lambda) = S \).

Any other HFLTS is formed with at least one linguistic term in \( S \).

**Example 1:** Let \( S \) be a linguistic term set \( S = \{ S_0: \text{nothing}, S_1: \text{very low}, S_2: \text{low}, S_3: \text{medium}, S_4: \text{high}, S_5: \text{very high}, S_6: \text{perfect} \} \) a different HFLTS might be

\[
H_0(\lambda) = \{ S_1: \text{very low}, S_2: \text{low}, S_3: \text{medium} \}
\]

\[
H_1(\lambda) = \{ S_3: \text{medium}, S_4: \text{high}, S_5: \text{very high}, S_6: \text{perfect} \}
\]

Once the concept of HFLTS has been defined, it is necessary to introduce the computation and operations that can be performed on them.

Let \( S \) be a linguistic term set, \( S = \{ S_0, \ldots, S_g \} \), and \( H_s, H_{s_1}, H_{s_2} \) be the three HFLTS.

**Definition 2:** The upper bound \( H_s^+ \) and lower bound \( H_s^- \) of the HFLTS, \( H_s \) are defined as

1) \( H_s^+ = \max(s_i) = s_j \), \( s_i \in H_s, s_i \leq s_j \) for all \( i \)

2) \( H_s^- = \min(s_i) = s_j \), \( s_i \in H_s, s_i \leq s_j \) for all \( i \)

**Definition 3:** The complement of HFLTS \( H_s \), is defined as

\( H_s = S - H_s = \{ s_i / s_i \in S, s_i \not\in H_s \} \).

**Definition 4:** The envelope of the HFLTS \( \text{env}(H_s) \), is a linguistic interval whose limits are obtained by means of upper bound (max) and lower bound (min). Hence \( \text{env}(H_s) = [H_s^-, H_s^+] \)

**Example 2:** Let \( S = \{ S_0: \text{nothing}, S_1: \text{very low}, S_2: \text{low}, S_3: \text{medium}, S_4: \text{high}, S_5: \text{very high}, S_6: \text{perfect} \} \) be a linguistic term set, and \( H_s = \{ \text{high, very high, perfect} \} \) be an HFLTS of \( S \), its envelope is

\( H_s^- = \{ \text{high, very high, perfect} \} = \text{high} \)

\( H_s^+ = \{ \text{high, very high, perfect} \} = \text{perfect} \)

\( \text{env}(H_s) = [\text{high, perfect}] \).
Table 1: Hesitant Fuzzy Linguistic Term Set (HFLTS)

<table>
<thead>
<tr>
<th>Data Set</th>
<th>$S={S_0$:nothing, $S_1$:very low, $S_2$:low, $S_3$:medium, $S_4$:high, $S_5$:very high, $S_6$:perfect$}$</th>
<th>$H_{s^+} = {\text{high, very high, perfect}} = \text{perfect}$</th>
<th>$H_{s^-} = {\text{high, very high, perfect}} = \text{high}$</th>
<th>$\text{env}(H_s) = [\text{high, perfect}]$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$D_1$</td>
<td>${0, 2, 3, 5, 7, 9, 6}$</td>
<td>${7.9, 6} = 6$</td>
<td>${7.9, 6} = 7$</td>
<td>${7, 6}$</td>
</tr>
<tr>
<td>$D_2$</td>
<td>${0.4, 7.9, 13.15, 12}$</td>
<td>${13.15, 12} = 12$</td>
<td>${13.15, 12} = 13$</td>
<td>${13, 12}$</td>
</tr>
<tr>
<td>$D_3$</td>
<td>${0, 1, 3, 6, 8, 10, 7}$</td>
<td>${8, 10, 7} = 7$</td>
<td>${8, 10, 7} = 8$</td>
<td>${8, 7}$</td>
</tr>
<tr>
<td>$D_4$</td>
<td>${0.6, 11, 17, 22, 29, 19}$</td>
<td>${22, 29, 19} = 19$</td>
<td>${22, 29, 19} = 22$</td>
<td>${22, 19}$</td>
</tr>
<tr>
<td>$D_5$</td>
<td>${0.3, 7, 8, 12, 13, 11}$</td>
<td>${12, 13, 11} = 11$</td>
<td>${12, 13, 11} = 12$</td>
<td>${12, 11}$</td>
</tr>
<tr>
<td>$D_6$</td>
<td>${0.9, 24, 29, 35, 40, 33}$</td>
<td>${35, 40, 33} = 33$</td>
<td>${35, 40, 33} = 35$</td>
<td>${35, 33}$</td>
</tr>
<tr>
<td>$D_7$</td>
<td>${0.7, 11, 16, 22, 27, 21}$</td>
<td>${22, 27, 21} = 21$</td>
<td>${22, 27, 21} = 22$</td>
<td>${22, 21}$</td>
</tr>
<tr>
<td>$D_8$</td>
<td>${0.5, 9, 14, 17, 23, 16}$</td>
<td>${17, 23, 16} = 16$</td>
<td>${17, 23, 16} = 17$</td>
<td>${17, 16}$</td>
</tr>
</tbody>
</table>

Table 2: Accuracy (in %) with Bag of Words Datasets, 20-news group Datasets and Legal Case Reports Datasets

<table>
<thead>
<tr>
<th>No. of Data Set</th>
<th>Bag of Word</th>
<th>20 News</th>
<th>Group</th>
<th>Legal Case</th>
<th>Report</th>
</tr>
</thead>
<tbody>
<tr>
<td>K-NN</td>
<td>SVM</td>
<td>K-NN</td>
<td>SVM</td>
<td>K-NN</td>
<td>HFLTS</td>
</tr>
<tr>
<td>50</td>
<td>0.84</td>
<td>0.97</td>
<td>0.87</td>
<td>0.89</td>
<td>0.83</td>
</tr>
<tr>
<td>100</td>
<td>0.86</td>
<td>0.94</td>
<td>0.86</td>
<td>0.71</td>
<td>0.82</td>
</tr>
<tr>
<td>200</td>
<td>0.82</td>
<td>0.92</td>
<td>0.85</td>
<td>0.73</td>
<td>0.81</td>
</tr>
<tr>
<td>350</td>
<td>0.80</td>
<td>0.93</td>
<td>0.79</td>
<td>0.76</td>
<td>0.90</td>
</tr>
<tr>
<td>500</td>
<td>0.77</td>
<td>0.89</td>
<td>0.73</td>
<td>0.78</td>
<td>0.84</td>
</tr>
<tr>
<td>650</td>
<td>0.78</td>
<td>0.88</td>
<td>0.76</td>
<td>0.79</td>
<td>0.75</td>
</tr>
<tr>
<td>800</td>
<td>0.73</td>
<td>0.86</td>
<td>0.72</td>
<td>0.72</td>
<td>0.73</td>
</tr>
<tr>
<td>1000</td>
<td>0.76</td>
<td>0.83</td>
<td>0.71</td>
<td>0.71</td>
<td>0.76</td>
</tr>
</tbody>
</table>

Table 3: F-measure value with Bag of Words datasets, 20-news group Datasets and Legal Case Reports Datasets

<table>
<thead>
<tr>
<th>No. of Data Set</th>
<th>Bag of Word</th>
<th>20 News</th>
<th>Group</th>
<th>Legal Case</th>
<th>Report</th>
</tr>
</thead>
<tbody>
<tr>
<td>K-NN</td>
<td>SVM</td>
<td>K-NN</td>
<td>SVM</td>
<td>K-NN</td>
<td>HFLTS</td>
</tr>
<tr>
<td>5</td>
<td>0.82</td>
<td>0.97</td>
<td>0.86</td>
<td>0.88</td>
<td>0.81</td>
</tr>
<tr>
<td>10</td>
<td>0.83</td>
<td>0.94</td>
<td>0.87</td>
<td>0.72</td>
<td>0.84</td>
</tr>
<tr>
<td>20</td>
<td>0.82</td>
<td>0.92</td>
<td>0.84</td>
<td>0.73</td>
<td>0.82</td>
</tr>
<tr>
<td>35</td>
<td>0.83</td>
<td>0.93</td>
<td>0.80</td>
<td>0.72</td>
<td>0.82</td>
</tr>
<tr>
<td>50</td>
<td>0.87</td>
<td>0.89</td>
<td>0.74</td>
<td>0.78</td>
<td>0.87</td>
</tr>
</tbody>
</table>
3. METHODOLOGY:
In Classification of document, different steps are used. The steps are shown in fig 2.

Fig 2: Hesitant Fuzzy Linguistic Term Set

4 EXPERIMENTAL RESULTS
In this Experiment we calculate Hesitant Fuzzy Linguistic Term Set (HFLTS) in Document dataset. Document Classification upper bound $H_+\$, envelope of the HFLTS and lower bound $H_-$ of the HFLTS HS are calculated which describe in Table 1 and Table 2 respectively. Table 3 to Table 5 describes classification accuracy for Bag of Words, 20-news group and Legal Case Reports Datasets respectively. Table 6 to Table 8 describes F-measure values for Bag of Words, 20-news group and Legal Case Reports Datasets respectively.

This Experiment shows, Hesitant Fuzzy Linguistic Term Set (HFLTS) based Document Classification is efficient and accurate compare to other Document Classification. From Fig.2 to Fig. 9, Hesitant Fuzzy Linguistic Term Set is described.
From Fig. 10 to Fig.11, Upper bound $H_{s^+}$, lower bound $H_{s^-}$ of the HFLTS are described respectively.

Fig. 12 to Fig.14 describes classification accuracy results for Bag of Words, 20-news group and Legal Case Reports Datasets respectively. Fig. 15 to Fig.17 describes F-measure values for Bag of Words, 20-news group and Legal Case Reports Datasets respectively.

$D_1$H+ upper bound, $D_1$H- lower bound are dataset 1, $D_2$H+ upper bound, $D_2$H- lower bound are dataset 2, $D_3$H+ upper bound, $D_3$H- lower bound are dataset 3, $D_4$H+ upper bound, $D_4$H- lower bound are dataset 4, $D_5$H+ upper bound, $D_5$H- lower bound are dataset 5, $D_6$H+ upper bound, $D_6$H- lower bound are Dataset 6, $D_7$H+ upper bound, $D_7$H- lower bound are dataset 7, $D_8$H+ upper bound, $D_8$H- lower bound are dataset 8 in graphical representation of Fig.2 to Fig.9 respectively.

![HFLTS](image1)

**Fig 2:** Hesitant Fuzzy Linguistic Term Set

![HFLTS](image2)

**Fig 3:** Hesitant Fuzzy Linguistic Term Set
Fig 4: Hesitant Fuzzy Linguistic Term Set

Fig 5: Hesitant Fuzzy Linguistic Term Set

Fig 6: Hesitant Fuzzy Linguistic Term Set
Fig 7: Hesitant Fuzzy Linguistic Term Set

Fig 8: Hesitant Fuzzy Linguistic Term Set

Fig 9: Hesitant Fuzzy Linguistic Term Set
Fig 10: Hesitant Fuzzy Linguistic Term Set

Fig 11: Hesitant Fuzzy Linguistic Term Set

Fig 12: Accuracy for Bag of Words datasets
**Fig 13:** Accuracy for 20-news group datasets

**Fig 14:** Accuracy for Legal Case Reports Datasets

**Fig 15:** F-measure for Bag of Words datasets
5. CONCLUSION:

As result & analysis shows that The Document Classification based on Hesitant Fuzzy Linguistic Term Set is efficient and the HFLTS classification has the potential to improve the classification accuracy.

REFERENCES


1673, 2006.


