Fuzzy Interval Numbers (FINs) techniques and applications

Theodoros Alevizos[1], Vassilis G. Kaburlasos[1], Stelios Papadakis[1], Christos Skourlas[2]
[1] Department of Industrial Informatics, T.E.I. of Kavala, Greece
Email: {alteo,vgkabs, spap}@teikav.edu.gr
[2] Department of Informatics, T.E.I. of Athens, Greece
Email: cskourlas@teiath.gr

Abstract

Fuzzy Interval Numbers (FINs) could be seen as a set of techniques applied in Fuzzy System applications. In this paper, definition, interpretation and a computation algorithm of FINs are briefly discussed. We also present a series of techniques based on Fuzzy Interval Numbers (FINs) to solve classification problems. Some examples showing the importance of these techniques for documents classification are also presented.

Keywords: Fuzzy Interval Number (FIN), Information Retrieval (IR), Fuzzy System applications, Classification.

1. Introduction

Fuzzy Interval Numbers (FINs) were introduced by Kaburlasos (see Petridis and Kaburlasos (2003), Kaburlasos (2004, 2006)) in Fuzzy System applications. The related theoretical (mathematical) background is based on the metric space of the generalized intervals. A FIN (see Figure 1) may be interpreted as a conventional Fuzzy Set; additional interpretations for a FIN are possible including a statistical interpretation.

Fuzzy Interval Numbers (FINs) and lattice algorithms have been employed in various real-world application including numeric and non-numeric data. Kaburlasos and Petridis (2000), Petridis and Kaburlasos (1998, 2000, 2001) presented FLN (Fuzzy Lattice Neurocomputing mainly for clustering. The most popular among the FLN models for clustering is σ-FLN. For example, Petrides and Kaburlasos (2000) describe the automated – electromechanical surgical mechatronics tool which is used to control penetration through soft tissues in the epidural punctury. In this framework a series of learning experiments for soft tissues recognition was carried out and the σ-FLN model and the Voting σ-FLNMAP algorithm were applied.

FLN algorithms were also used in stapedotomy surgery (Kaburlasos et al 1997), and prediction of ozone concentration by classification, based on meteorological and air quality data (Athanasiadis and Mitkas 2004).

Petridis et al. (2001), Kaburlasos et al (2002), Petridis and Kaburlasos (2003) reported a best sugar prediction accuracy using Fuzzy Interval Numbers and the FINkNN classifier and a population (measurements) of production and meteorological data. Kaburlasos et al (2005) used FINs for representing geometric and other fertilizer granule features. This type of modelling was used to cover needs of the Greek Fertilizer Industry. Marinagi et al (2006) propose and discuss the use of FINs classifier to handle problems of Cross Language Information Retrieval.

In section 2 a brief introduction to definition and construction of FINs is given. In section 3 the FINs (vectors) similarity is defined. Section 4 presents Fuzzy Interval Numbers and documents' classification, examples of our experimentation and the evaluation of FIN - techniques. Conclusions are presented in section 5.

2. Definition and construction of FINs

FIN could be regarded as an abstract "mathematical object" and can have various interpretations and uses.

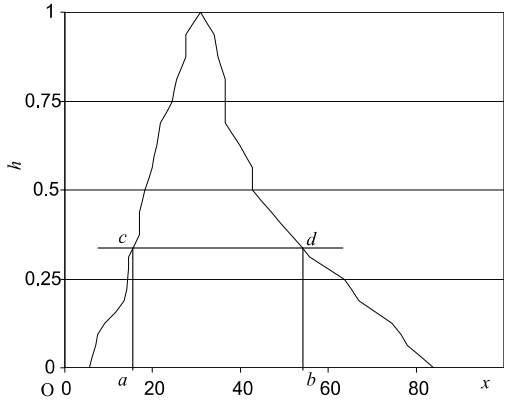

Figure 1. A FIN constructed by the CALFIN algorithm.

Given a population (a vector) $x = [x_1,x_2,\ldots,x_N]$ of measurements (real numbers), sorted in ascending order. The dimension of a vector x is denoted by dim(x) e.g. dim([2,-1]) = 2, dim([-3,4,0,-1,7]) = 5. The median(x) of the vector $x = [x_1,x_2,\ldots,x_N]$ is defined to be a number such that half of the N numbers $x_1,x_2,\ldots,x_N$ are smaller than median(x) and the other half are larger than median(x); for instance, $\text{median}([x_1,x_2,x_3]) = x_2$, with $x_1 < x_2 < x_3$, whereas $\text{median}([x_1,x_2,x_3,x_4]) = (x_2 + x_3)/2$, with $x_1 < x_2 < x_3 < x_4$. A FIN can be computed for the population (the vector x) by applying the following CALFIN algorithm (see Kaburlasos, 2006).

Algorithm CALFIN // Calculate FIN

Let x be a vector of measurements (real numbers).
Sort vector x incrementally.
Initially vector pts is empty.
function calfin(x) {
while (dim(x) ≠ 1)
medi:= median(x)
insert medi in vector pts
x_left:= elements in vector x less-than number median(x)
x_right:= elements in vector x larger-than number median(x)
calfin(x_left)
calfin(x_right)
endwhile
} //function calfin(x)
Sort vector pts incrementally.
Store in vector val, dim(pts)/2 numbers from 0 up to 1 in steps of 2/dim(pts) followed by another dim(pts)/2 numbers from 1 down to 0 in steps of 2/dim(pts).

For example, let us consider the following vector x=(4.333, 4.334, 4.335, 4.667, 4.668, 5, 5.25, 5.251, 5.252, **5.5**, 5.501, 5.75, 6, 8, 8.001, 8.5, 8.501, 8.502, 9)
It follows that median(x) = 5.5 and,
vector(x1) = x_left = (4.333, 4.334, 4.335, 4.667, 4.668, 5, 5.25, 5.251, 5.252),
vector(x2) = x_right=(5.501, 5.75, 6, 8, 8.001, 8.5, 8.501, 8.502, 9)
Then, median(x1_left) = median (x_left)= 4.668 and
vector(x11)=x1_left = (4.333, 4.334,4.335, 4.667),
vector(x12)=x2_right = (5, 5.25, 5.251, 5.252)
Then median(x2)=median(x2_right) = 8.001 and
vector(x21)= x2_left = (5.501, 5.75, 6, 8)
vector(x22)= x2_right= (8.5, 8.501, 8.502, 9)
etc

The above procedure is repeated recursively $\log_2 N$ times, until "half vectors" are computed including a single number; the latter numbers are, by definition (construction), median numbers. The computed median values are stored (sorted) in vector pts, whose entries constitute the abscissae of a positive FIN's membership function; the corresponding ordinate values are computed in vector val.

Eventually Algorithm CALFIN computes two vectors, i.e. pts and val, where vector val includes the degrees of (fuzzy) membership of the corresponding real numbers in vector pts. Note that algorithm CALFIN produces a positive FIN with a membership function $\mu(x)$ such that $\mu(x)=1$ for exactly one number x (Kaburlasos, 2006).
In the figure 1 a FIN is constructed by the CALFIN algorithm. Any "cut" at a given height $h \in (0,1]$ defines a generalized interval, denoted by F(h) or $[c, d]^h$. In the area [cdba], which is the support (F(0.25)), there are about 75% of the values.

In the figure 2 two documents (vectors) are illustrated (plotted) as two Fuzzy Interval Numbers. Each value on the term axis represents a term (stem). Any "cut" at a given height $h \in (0,1]$ defines two generalized intervals, denoted by $[a, c]^h$, $[b, d]^h$. In our case the generalized intervals are positive and intersecting. As we can see below, a bell-shaped mass function (see figure 2) is used for the calculations of the distance between two FINs (or the similarity between two documents which are represented by their FINs).

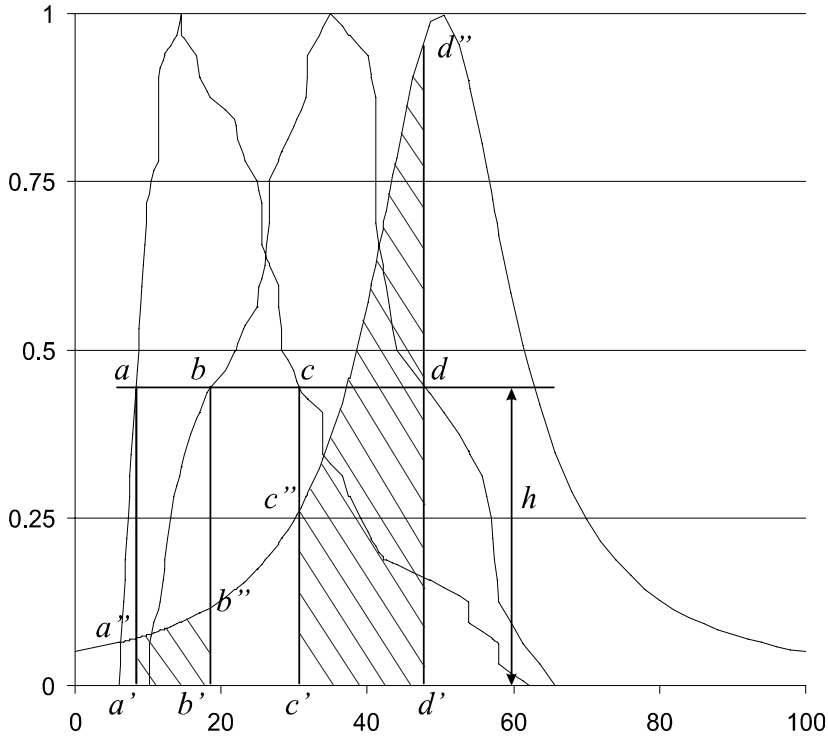

Figure 2. Two documents (vectors) illustrated as two Fuzzy Interval Numbers. Each value on the term axis represents a term (stem). A bell-shaped mass function is used for the calculation of the distance between the documents.

3. Fuzzy Interval Numbers and FINs (vectors) similarity

The concept of the Generalized Intervals is used for introducing a metric into the lattice of the Fuzzy Interval Numbers (FINs) [Kaburlasos 2006]. The interpretation of a generalized interval depends on an application; for instance if a feature is present in a document it could be indicated by a positive generalized interval [Marinagi 2006].

The area "under" a generalized interval is a real number which could be calculated. We can define a metric and an inclusion measure function in the set (lattice) of the generalized intervals $\mathbf{M}_h$ [Kaburlasos 2006].

A positive generalized interval of height h∈(0,1], is denoted by $[x_1, x_2]^h$ (or $\mu_{[x_1,x_2]^h}$) and is defined as a map $[x_1, x_2]^h : R \longrightarrow \{0, h\}$, given by:

$[x_1, x_2]^h = h$ ($x_1 \leq x \leq x_2$) and $[x_1, x_2]^h = 0$, otherwise ($x_1 \leq x_2$).

A negative generalized interval of height h∈(0,1] is defined as a map $[x_1, x_2]^h : R \longrightarrow \{0, -h\}$, given by:

$[x_1, x_2]^h = -h$ ($x_1 \geq x \geq x_2$) and $[x_1, x_2]^h = 0$, otherwise ($x_1 > x_2$).

The set of all positive generalized intervals of height h is denoted by $\mathbf{M}_+^h$, the set of all negative generalized intervals by $\mathbf{M}_-^h$, and the set of all generalized intervals by $\mathbf{M}^h$.

3.1 FIN metrics

Let m_h: $\mathbf{R} \rightarrow \mathbf{R}^+$ be a positive real function – a mass function – for h ∈ (0,1] (could be independent of h) and $f_h(x) = \int_0^x m_h(t)dt$.

Function f_h is strictly increasing.

The real function v_h: $\mathbf{M}_+^h \rightarrow \mathbf{R}$, given by $v_h([a,b]^h) = f_h(b) - f_h(a)$, is a positive valuation function in the set of positive generalized intervals of height h.

Therefore, $d_h([a,b]^h, [c,d]^h) = [f_h(a \vee c) - f_h(a \wedge c)] + [f_h(b \vee d) - f_h(b \wedge d)]$,

where $a \wedge c = \min\{a,c\}$ and $a \vee c = \max\{a,c\}$,

and $d_h([a,b]^h,[c,d]^h)$ is a metric between the two generalized intervals $[a,b]^h$ and $[c,d]^h$.

Given two positive FINs F_1 and F_2, then

$$d(F_1,F_2) = c\int_0^1 d_h(F_1(h), F_2(h))dh$$

where c is an user-defined positive constant, is a metric ~~distance~~ (for a proof see (Kaburlazos, 2004))

3.2 Calculation of the distance between documents

The FIN distance is used instead of the similarity measure between documents: the smaller the distance the more similar the documents.

In the case of documents and for the distance calculations a bell-shaped mass function could be selected:

$$m_h(t) = \frac{\rho + (1-\rho)h}{1 + \frac{1-z}{z}\left(\frac{t}{maxctf/\nu} - 1\right)^2}[\sigma + (1-\sigma)t]$$

The positive real numbers ρ, σ, z, t, v are parameters; maxctf is the maximum value of all the collection term frequencies. Figure 2 illustrates how we try to calculate the distance of two FINs F_1 and F_2 (representing two documents) using the mass function. The points *a*, *b*, *c*, *d* are used to define the distance, at height *h,*

$d_h(F_1(h),F_2(h)) = d_h([a,b]^h,[c,d]^h)$;

$d_h(F_1(h),F_2(h))$ equals the sum of areas of the shaded regions.

The distance of the two FINs at the height h is given by the sum of areas of the shaded regions.
The distance between the two FINs is calculated using the definite integral of the distance at height h from h=0 to 1:

$$\text{Distance} = \int_0^1 (\left| \int_a^b f(t)dt \right| + \left| \int_c^d f(t)dt \right|)dh$$

4. Fuzzy Interval Numbers and documents' classification

Fuzzy (set) techniques were proposed for Information Retrieval (IR) applications many years ago (Radecki, 1979), (Kraft, 1993), mainly for modeling.

The basic features of a FINs method for documents' classification are the following:

1) Documents are represented as FINs; a FIN resembles a probability distribution.

2) The FIN representation of documents is based on the use of the collection term frequency as the term identifier.

3) The use of FIN distance instead of a similarity measure.

4.1 Examples of successful classification

We form various samples using bibliographic descriptions extracted from the (bibliographic) databases of the Greek National Documentation Centre (NDC) and conduct several experiments. The rationale for selecting such bibliographic records is the following:

Each bibliographic record (text, document) is publicly available in various formats (e.g. text, MARC XML). There are various databases that cover different research topics (e.g. medical bibliography, dissertations). Each record is described by the title(s), the abstract(s), the key-phrases etc. All these fields of the description are bilingual: text in Greek and English (or other language e.g. French). Unfortunately,

there are bibliographic descriptions that are not complete e.g. abstracts are not included in some cases.

Every sample was constructed as the union of the retrieved documents by simple queries (searches) based on a single search term.

Sample 1: Medical bibliographic records

The records of the NDC were searched using the search terms «παιδιατρική» (extracted 9 records - documents containing the search term in their key-phrases: 782, 3282, 3371, 6570, 8303, 8421, 9202, 10481, 10711), "pediatrics" (extracted 5 records - documents not containing the search term in their key-phrases: 6525, 11102, 11931, 12532, 12498), and "mastectomy" (25 records were extracted: 183, 185, 1289, 1625, 1627, 1978, 1990, 2118, 3013, 4958, 5512, 6538, 6540, 7977, 8040, 8252, 8755, 9434, 9510, 10734, 10871, 11136, 11192, 12499, 14183. 16 records contain the search term in their keywords).

All the records without abstract were deleted and then all the documents that have a search term in their keywords formed two classes: Pediatrics-class, Mastectomy-class. Then the rest of the documents were classified successfully using the FIN-based similarity (see Table 1 for some calculations)

Table 1. *Medical documents' classification*

record	Average distance from the Pediatrics – class	Average distance from the Mastectomy - class	Comments
6525	0.2048	0.3769	Correct
11102	0.2288	0.3836	Correct
11931	0.2129	0.4252	Correct
12532	0.1891	0.3116	Correct
12948	0.2930	0.4910	Correct

Sample2: Dissertations.

The retrieved documents were extracted from a bibliographic database which is part of the Greek National archive of Dissertations. Our sample was constructed as the union of the retrieved documents by two simple queries (searches):

Using the (search) term "Natural Language Processing" we retrieved seven documents. Five documents contain the search term in their key-phrases and the other documents contain the term in other fields e.g. the abstract. Using the search term "Information Retrieval" we retrieved twelve documents and six of them contain the search term in their key-phrases.

The results of the two searches formed two classes of documents (dissertations):

{8011, 8432, 8433, 8728, 11137, 11646, 12648}

{909, 2792, 3015, 3171, 7781, 8553, 8556, 11143, 12197, 12424, 13239, 13290}

Then, we formed two classes (sub-collections) including only the documents that contain the search term in their key-phrases.:

NLP-class = {8011, 8432, 8433, 11137, 12648}

IR-class = {3171, 8553, 8556, 11143, 12424, 13290}

We measured the average FIN-distance of the other documents from the two classes using the proposed FIN-based technique and the documents were classified correctly as you can see in the table 2.

Table 2. *Dissertations' classification*

Document	IR-collection	NLP-Collection	Comments on the technique used for calculating the average distance
8728	0.3228	0.1865	Correct
11646	0.6788	0.2606	Correct
8728Gre	0.5889	0.1507	Correct
8728Eng	0.7551	0.1503	Correct
3171Gre	0.1088	0.1451	Use Weights or Centroid
3171Eng	0.2904	0.2922	Use Weights or Centroid
8553Gre	0.1773	0.2220	Use Weights or Centroid
8553Eng	0.2141	0.3833	Correct
8553Eng	<=0.0680	0.0776	Use only the closer distances for x (= 4 or 5) elements of the class

When the documents were restricted only in the English part (e.g. title and abstract only in English) the documents' length was reduced and some documents were erroneously classified. This was an indication that our technique is more appropriate in the case of "large" documents. Hence, in some cases there was a need to "correct" the classification errors (and in general improve our results) using weights (or penalties or calculation of the average distance only from the top-x distances). Significant improvement could be also mentioned in the case of using the classes' centroids.

5. Conclusions and future activities

We have concluded that the FIN-based calculation of the similarity between measurements (vectors) is a novel method for solving various problems. Especially, such a method seems to be promising in the case of documents classification. We verified in our experiments that if we focus on the documents that are related to different search terms then we can easily apply FIN-based techniques and calculate correctly the distance (similarity) between them. If the length of the documents is greater then the results of the method are much better.

Documents' classification can include various strategies:

You can use a training set and form classes ("sub-collections") of the sample and then calculate the FIN-distance of the "unclassified" documents from these classes. This distance could be seen as an average distance of every unclassified document from all the elements of the class. You can also consider only the (top-x) documents which are closer to the unclassified one. Then you can calculate the average distance of the unclassified document from these (top-x) documents.

Classification is more accurate if you use weights for the search terms that are contained in the retrieved documents and/or penalty (negative weight) in the case that the search term is not included.

A strategy related to the specific sample of classified and unclassified documents could be defined and you can combine average distance calculation and (if it is necessary) weights and / or penalty. As an example, if the general (average) distance of a document from a class (which is calculated from the distances of the document from all the classified documents of the class) and the partial one (which is calculated from a number of top-x distances) are not "consistent" you can use weights following the appropriate technique.

It is worthwhile to mention that if you use centroids, one for each class, you can improve the accuracy of documents' classification, in general.

At present, experiments with promising results are been conducted mainly with other bilingual samples extracted from the databases of the Greek NDC and three of the standard monolingual collections, namely MED, CACM and WSJ. These standard collections are of interest because they have relatively long queries.

6. References

Radecki,T (1979), "Fuzzy Set Theoretical Approach to Document Retrieval" in Information Processing and Management, v.15, Pergammon-Press 1979.

Kraft, D.H. and D.A. Buell (1993), "Fuzzy Set and Generalized Boolean Retrieval Systems" in Readings in Fuzzy Sets for Intelligent Systems, D. Dubius, H.Prade, R.R. Yager (eds).

Kaburlazos V. G. (2006), Towards a Unified Modeling and Knowledge – Representation based on Lattice Theory - Computational Intelligence and Soft Computing Applications, Springer-Verlag, Summer 2006

Kaburlasos, V.G. (2004), "Fuzzy Interval Numbers (FINs): Lattice Theoretic Tools for Improving Prediction of Sugar Production from Populations of Measurements," IEEE Trans. on Man, Machine and Cybernetics – Part B, vol. 34, no 2, pp. 1017-1030, 2004.

Petridis, V. and V.G. Kaburlasos (2003), "FINkNN: A Fuzzy Interval Number k-Near-est Neighbor Classifier for prediction of sugar production from populations of samples," Journal of Machine Learning Research, vol. 4 (Apr), pp. 17-37, 2003

Kaburlasos and Petridis (2000), Fuzzy Lattice Neurocomputing models, Neural Networks, 13(10), 1145-1170, 2000.

Petridis and Kaburlasos (1998), Fuzzy lattice neural network (FLNN): A hybrid model for learning, IEEE Trans. Neural Networks, 9(5), 877-890, 1998.

Petridis and Kaburlasos (2000), An intelligent mechatronics solution for automated tool guidance in the epidural surgical procedure, Proc. 7th Annual conf. Mechatronics and Machine Vision in Practice, pp 201-206, 2000.

Petridis and Kaburlasos (2001), Clustering and classification in structured data domains using Fuzzy Lattice Neurocomputing, IEEE Trans. Knowledge Data Engineering, 13(2), 245-260, 2001

Kaburlasos et al (1997), Automatic detection of bine breakthrough in orthopedics by fuzzy lattice reasoning: The case of drilling in the osteosynthesis of long bones, Proc. Mechatronics Computer systems for Perception and Action, pp 33-40, 1997.

Athanassiadis and Mitkas (2003), Applying machine learning techniques on air quality data for real-time decision support, Proc. Intl. NAISO Symposium on Information Technologies in Environmental Engineering, 2003.

Kaburlasos V.G., Spais V, Petridis V, Petrou L, Kazarlis S, Maslaris N, and Kallinakis A (2002), Intelligent clustering techniques for prediction of sugar production, Mathematics and Computers in Simulation, 60(3-5), 159-168, 2002

Kaburlasos V.G. Papadakis S. (2005) granular Self Organizing Map (grSOM) neural network for industrial quality control, Proc of SPIE, Mathematical Methods in Pattern and Image Analysis, 2005

Marinagi, Alevisos, Kaburlasos, Skourlas (2006), Fuzzy Interval Number (FIN) Techniques for Cross Language Information Retrieval, Proc. 8th ICEIS, 2006