Knowledge-Based Systems 70 (2014) 335–344

Contents lists available at ScienceDirect

Knowledge-Based Systems

journal homepage: www.elsevier.com/locate/knosys

A penalty-based aggregation operator for non-convex intervals

Gleb Beliakov, Simon James *

Deakin University, School of Information Technology, 221 Burwood Hwy, Burwood, Victoria 3125, Australia

ARTICLE INFO

Article history:
Received 3 September 2013
Received in revised form 22 May 2014
Accepted 17 July 2014
Available online 29 July 2014

Keywords:
Aggregation functions
Penalty-based functions
Interval-valued fuzzy sets
Hesitant fuzzy sets
Averaging operators

ABSTRACT

In the case of real-valued inputs, averaging aggregation functions have been studied extensively with results arising in fields including probability and statistics, fuzzy decision-making, and various sciences. Although much of the behavior of aggregation functions when combining standard fuzzy membership values is well established, extensions to interval-valued fuzzy sets, hesitant fuzzy sets, and other new domains pose a number of difficulties. The aggregation of non-convex or discontinuous intervals is usually approached in line with the extension principle, i.e. by aggregating all real-valued input vectors lying within the interval boundaries and taking the union as the final output. Although this is consistent with the aggregation of convex interval inputs, in the non-convex case such operators are not idempotent and may result in outputs which do not faithfully summarize or represent the set of inputs. After giving an overview of the treatment of non-convex intervals and their associated interpretations, we propose a novel extension of the arithmetic mean based on penalty functions that provides a representative output and satisfies idempotency.

1. Introduction

The arithmetic mean is the standard "go to" operator employed in various sciences, statistics, economics and fuzzy decision making for aggregating a set of inputs into a single representative value. For example, a number of small errors may arise naturally when we conduct repeated experiments or in our data collection, so we use the arithmetic mean to average the results, providing a reasonable estimate of what might be the *true* value. In fuzzy decision making, the arithmetic mean of the expert evaluations or membership values can be used to compare potential alternatives, allowing us to choose the best overall option. On the other hand, we may simply be interested in a summary statistic that tells us in some way what is normal or expected for a particular set of inputs, e.g. to describe a population in terms of the average life expectancy.

More broadly, the arithmetic mean is one example of an averaging aggregation function [5,14,23]. Aggregation functions have been studied for various practical applications and a number of alternatives to the arithmetic mean have been proposed that may perform more reliably for certain types of data. In the face of uncertainty pertaining to the inputs, either arising from linguistic descriptions or data collection methods, the need has also been identified to extend aggregation functions to deal with inputs expressed as intervals [11], pairs of positive and negative information such as Atanassov orthopairs [1] or other multiple-valued inputs [2,3].

Although a number of results have been established for these data types (especially in the field of fuzzy sets aggregation), more recently some researchers have tackled the problem of aggregating inputs provided as non-convex or discontinuous intervals, i.e. intervals that contain gaps or are comprised of a sequence of disjoint intervals. In statistics research, such inputs can occur through censoring where the data cannot be observed over particular intervals. In probability theory, the study of random sets also gives rise to non-convex sets and the need to calculate their expectation (see [20] for a detailed overview). For examples in the fuzzy research community, we can mention the hesitant fuzzy sets of Torra and Narukawa [22,24] – where the input usually denotes a discrete set of possible evaluations between 0 and 1, the generalized gray numbers of Yang and John [29,30] – where an input is known to lie within a potentially non-convex range of values, and the discontinuous intervals of Wagner et al. [26].

These are inputs of the form

$$A_i = \bigcup_{j=1}^{m_i} [a_{i_j}^-, a_{i_j}^+], \tag{1}$$

where $[a_{i_j}^-, a_{i_j}^+]$ denotes the j-th interval with $a_{i_j}^+ < a_{i_{j+1}}^-$, $j = 1, 2, \dots, m_i - 1$. It may also be convenient to represent

* Corresponding author. Tel.: +61 3 9251 7481.
E-mail addresses: gleb@deakin.edu.au.au (G. Beliakov), sjames@deakin.edu.au (S. James).

http://dx.doi.org/10.1016/j.knosys.2014.07.011
0950-7051/

such intervals as a sequence of intervals (as in [16]), i.e. for $A_i = [a^-_{i_1}, a^+_{i_1}] \cup, \ldots, \cup [a^-_{i_{m_i}}, a^+_{i_{m_i}}]$ we will simply write $A_i = \langle [a^-_{i_1}, a^+_{i_1}], \ldots, [a^-_{i_{m_i}}, a^+_{i_{m_i}}] \rangle$.

We can consider the following situations where it may be useful to work with non-convex intervals.

Example 1 (*Uncertainty with travel times [30]*). Two trains are scheduled to depart 5 min apart however both could be up to 2 min late. The earlier train is sometimes full and it takes 3 min to travel to the next station. The time it will take a passenger (arriving in time for the first train) to reach the next station can be represented with the input $\langle [3,5],[8,10] \rangle$.

Example 2 (*Species population recovery intervals [21]*). An ecology expert is asked to provide her estimation of when a species will reach healthy population levels following a forest fire. A species may increase in population immediately following the fire, then decrease as other species start to recover, before increasing again to its pre-disturbance levels. The expert can use the non-convex interval $\langle [1,2],[8,20] \rangle$ to indicate that the population is predicted to be at healthy levels 1–2 years and 8–20 years after the fire, but below the threshold at other times.

Example 3 (*Analysis of recurring health problems [15]*). After a patient is treated and released from hospital, they may make a full recovery or sometimes their health will deteriorate and they will be readmitted. In some cases, the patient may be readmitted a number of times. In order to investigate contributing factors, experts represent each patient's time in hospital with non-convex intervals, e.g. an input $\langle [12,13],[16,18],[24,27] \rangle$ would indicate a patient was readmitted for 3 periods after 12, 16 and 24 months, and stayed in hospital for various lengths of time.

In each of the examples above, the inputs represent temporal data [16], with an event (or the uncertainty pertaining to an event) taking duration over discontinuous time periods. However there may also be cases where we need to aggregate non-convex intervals that represent spatial data, e.g. in the fusion of sensor readings observing a non-continuous space, measurements that are uncertain because they lie outside an observable range (censored data), or even evaluations in fuzzy decision making [24]. The authors in [26] also note example applications in forensic science, hazard detection, agreement-based modeling and computing with respect to linguistic descriptions.

In research areas such as statistics, a common approach for handling either standard interval or non-convex interval inputs is to represent them with single values, e.g. a mid-point or the most probable value. While this may be effective under certain conditions, working with the inputs in their original form can allow for robust analyses and inferences which are free of assumptions pertaining to the source of uncertainty [11].

The extension of averaging aggregation functions to non-convex inputs in the fuzzy domain has thus far been approached in a manner consistent with the extension principles applied for fuzzy operations (e.g. in [10] for operations on fuzzy numbers) and interval arithmetic [11]. All possible real-valued input vectors $\mathbf{x}$ with $x_i \in A_i \ \forall i$ are aggregated and the union of these aggregated values is taken as the output. We contend that while this approach is suitable for most situations when the intervals are convex, the non-convex case presents two unique problems:

1. These resulting aggregation functions are not idempotent, i.e. it does not necessarily hold that $f(t,t,\ldots,t) = t$ if t is a non-convex interval. Idempotency is a key property for averaging functions when it is desired that the output gives a representative or typical value. If we wanted to consider the "average" evaluation from 10 ecology experts in Example 2 above and all 10 provided the same interval $\langle [1,2],[8,20] \rangle$, then we would expect the same non-convex interval to be returned as the output.
2. As the number of inputs grows large, aggregating non-convex inputs in this fashion converges towards aggregating their convex hulls or envelopes (the intervals defined by the lower and upper bounds). This raises the question of whether anything is gained by using non-convex intervals to represent the uncertainty of the model in the first place.

In this article, we approach the problem of aggregating non-convex intervals in the framework of penalty-based functions. We show that existing methods can be recovered with the choice of various penalties and then propose a new penalty which results in an extension of the arithmetic mean which is idempotent and can more faithfully represent a "typical input".

The article will be structured as follows. In Section 2 we will give an overview of the concepts that underlie the proposed operators. In particular, we look at aggregation functions, penalty-based methods for constructing them, and various types of inputs for which they have been defined. In Section 3, we recall the definitions of functions which have been used in various settings to aggregate non-convex interval valued inputs, noting their relationship to penalty functions defined for intervals. We then consider the problem of defining penalties between non-convex intervals in Sections 4 and 5, we propose a new penalty for non-convex inputs and define our new operator. We present some numeric examples in Section 6 to help illustrate differences between the existing and proposed methods, before discussing some other potential approaches in Section 7 and concluding in Section 8.

2. Preliminaries

We will firstly provide the basic definitions relating to aggregation functions and show how they can be defined with respect to penalty functions. We then give an overview of various input types which have extended the use of real inputs to incorporate uncertainty into decision processes and modeling applications.

We will consider aggregation functions (see [5,14,23]) defined over the unit interval.

Definition 1 (*Aggregation function*). An aggregation function $f : [0,1]^n \to [0,1]$ is a function non-decreasing in each argument and satisfying $f(0,\ldots,0) = 0$ and $f(1,\ldots,1) = 1$.

The monotonicity of aggregation functions is important when used for decision making to ensure that an increase to one of the criteria should not result in a decrease in the overall evaluation. Here we are interested particularly in averaging aggregation functions.

Definition 2 (*Averaging aggregation function*). An aggregation function f is considered to be *averaging* where for $\mathbf{x} \in [0,1]^n$,

$$\min(\mathbf{x}) \leqslant f(\mathbf{x}) \leqslant \max(\mathbf{x}).$$

Due to the monotonicity of aggregation functions, averaging behavior is equivalent to *idempotency*, i.e. $f(t,t,\ldots,t) = t$.

Typical examples include the arithmetic mean (sometimes referred to simply as "the average") and the median. For an input vector $\mathbf{x}$ consisting of n values, the arithmetic mean $AM : \mathbf{x} \in [0,1]^n \to [0,1]$ is given by

$$AM(\mathbf{x}) = \sum_{i=1}^{n} \frac{1}{n} x_i. \tag{2}$$