Information Sciences 281 (2014) 1–7

Contents lists available at ScienceDirect

Information Sciences

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

On some properties of weighted averaging with variable weights

Gleb Beliakov *, Tim Wilkin

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

ARTICLE INFO

Article history:
Received 12 November 2013
Received in revised form 2 May 2014
Accepted 8 May 2014
Available online 16 May 2014

Keywords:
Aggregation function
Mean
Density based averaging
Monotonicity
Weak monotonicity

ABSTRACT

Density-based means have been recently proposed as a method for dealing with outliers in the stream processing of data. Derived from a weighted arithmetic mean with variable weights that depend on the location of all data samples, these functions are not monotonic and hence cannot be classified as aggregation functions. In this article we establish the weak monotonicity of this class of averaging functions and use this to establish robust generalisations of these means. Specifically, we find that as proposed, the density based means are only robust to isolated outliers. However, by using penalty based formalisms of averaging functions and applying more sophisticated and robust density estimators, we are able to define a broader family of density based means that are more effective at filtering both isolated and clustered outliers.

1. Introduction

Several novel data fusion operators have recently been proposed where the weights applied to the inputs are not constant [1,12,13]. Some of these operators are derived from a general class of averaging functions known as Bajraktarevic means, which generalise quasi-arithmetic means [3,4] by applying weights that depend on the inputs. A specific example are mixture operators studied in [10–12,15]. Statistically grounded logic operators [13] are based on the minimisation of a penalty (the sum of absolute differences between the inputs and the output), which is weighted by a function of the inputs. The most recent proposal is density-based averaging [1], which was derived from a weighted arithmetic mean, but involves variable weights which depend on the density of inputs. Here the weighting functions depend not only on their respective inputs, but on all the inputs and hence they are a more general construction than Bajraktarevic means.

In general, averaging functions with weights that depend on the inputs $\mathbf{x} \in \mathbb{I}^n \subseteq \mathbb{R}^n$ are not monotone in $\mathbf{x}$ and hence cannot be classified as aggregation functions [3]. Nevertheless, due to the broad interest in such functions and their potential to improve performance in real world applications, it is worthwhile to establish the mathematical properties that these functions possess. Specifically we are interested to show whether or not there is a less restrictive type of monotonicity which is meaningful in terms of the applications. As our study shows, there is a property which we call *weak monotonicity* which is useful for understanding these means, for developing generalisations of this family of functions and for improving the interpretation of results obtained when applying them.

An important driver in the development of means with variable weights is the need to deal with outliers, which exist in most real data sets. Outliers are data that stand apart from the main trend of inputs and can be due either to errors in

* Corresponding author. Tel.: +61 392517475.
E-mail addresses: gleb@deakin.edu.au (G. Beliakov), tim.wilkin@deakin.edu.au (T. Wilkin).

http://dx.doi.org/10.1016/j.ins.2014.05.011
0020-0255/© 2014 Elsevier Inc. All rights reserved.

measurement or recording, or unaccounted for phenomena. In the first case outliers need to be eliminated so as not to contaminate the data, its interpretation or subsequent analysis. In the second case the outliers are themselves the main interest. We argue that most monotone functions are not suitable for accounting for outliers (a notable exception is the median), because monotonicity in all arguments implies that the output is always affected by variation in any input, whether or not it is far from the main trend. Ideally we should weigh the contribution of each input to the average according to how far away this input is from the main trend (or majority) of the inputs, as has been proposed in the density based averaging concept [1]. However, once we introduce input-dependent weights and are able to eliminate the outliers, we lose monotonicity, since the product of a decreasing weight and increasing input (moving away from the remaining inputs) is not a monotone function [11].

In this article we examine mathematical properties of some averages with variable weights using the concept of weak monotonicity, introduced in [22]. We will show that density based averages are weakly monotone and bounds preserving. However, we will show that the claim made in [1] regarding the ability to effectively filter outliers holds only in special cases. We subsequently generalise density based means in two directions. First, we modify the expression into its penalty based representation and second, we consider more sophisticated and robust density estimators. The resulting families of density based averages more effectively filter multiple outliers from the data.

The remainder of the short paper is structured as follows. Section 2 provides preliminaries regarding means and discusses the penalty based representation of averaging functions. Section 3 presents the property of weak monotonicity and density-based averaging is discussed in detail in Section 4, where our generalisations are presented. Section 5 concludes this work.

2. Means with variable weights

2.1. Aggregation functions

In this article we make use of the following notations and assumptions. Without loss of generality we assume that the domain of interest is any closed, non-empty interval $\mathbb{I} = [a, b] \subseteq \overline{\mathbb{R}} = [-\infty, \infty]$ and that tuples in $\mathbb{I}^n$ are defined as $\mathbf{x} = (x_{i,n} | n \in \mathbb{N}, i \in \{1, \ldots, n\})$. We write x_i as the shorthand for $x_{i,n}$ such that it is implicit that $i \in \{1, \ldots, n\}$. Furthermore, $\mathbb{I}^n$ is ordered such that for $\mathbf{x}, \mathbf{y} \in \mathbb{I}^n$, $\mathbf{x} \leqslant \mathbf{y}$ implies that each component of $\mathbf{x}$ is no greater than the corresponding component of $\mathbf{y}$, i.e., $x_i \leqslant y_i \forall i \in \{1, 2, \ldots, n\}$. Unless otherwise stated, a constant vector given as $\mathbf{c}$ is taken to mean $\mathbf{c} = c(\underbrace{1, 1, \ldots, 1}_{n\ \textbf{times}}) = c\mathbf{1}c \in \mathbb{R}$, where n is implicit within the context of use. We will make use of the common shorthand notation for a sorted vector, being $\mathbf{x}_{()} = \{x_{(1)}, x_{(2)}, \ldots, x_{(n)}\}$. In such cases the ordering will be stated explicitly and then $x_{(k)}$ represents the kth largest or smallest element of $\mathbf{x}$ accordingly.

Consider now the following definitions:

Definition 1. A function $F : \mathbb{I}^n \to \bar{\mathbb{R}}$ is **monotonic** (non-decreasing) if and only if, $\forall \mathbf{x}, \mathbf{y} \in \mathbb{I}^n, \mathbf{x} \leqslant \mathbf{y}$ then $F(\mathbf{x}) \leqslant F(\mathbf{y})$.

Definition 2. A function $F : \mathbb{I}^n \to \mathbb{I}$ is an **aggregation function** in $\mathbb{I}^n$ if and only if F is monotonic non-decreasing in $\mathbb{I}$ and $F(\mathbf{a}) = a, F(\mathbf{b}) = b$.

Thus the two fundamental properties defining an aggregation function are monotonicity with respect to all arguments and bounds preservation. Further properties of aggregation functions relevant within this article are:

Definition 3. A function F is called **idempotent** if for every input $\mathbf{x} = (t, t, \ldots, t),\ t \in \mathbb{I}$ the output is $F(\mathbf{x}) = t$.

Definition 4. A function F has **averaging behaviour** (or is averaging) if for every $\mathbf{x}$ it is bounded by $\min(\mathbf{x}) \leqslant F(\mathbf{x}) \leqslant \max(\mathbf{x})$.

Due to monotonicity, aggregation functions that have averaging behaviour are idempotent and vice versa. Of particular relevance to us is the notion of shift-invariance. A constant change in every input should result in a corresponding change of the output.

Definition 5. A function $F : \mathbb{I}^n \to \mathbb{I}$ is **shift-invariant** if $F(\mathbf{x} + c\mathbf{1}) = F(\mathbf{x}) + c$ whenever $\mathbf{x}, \mathbf{x} + c\mathbf{1} \in \mathbb{I}^n$ and $F(\mathbf{x}) + c \in \mathbb{I}$.

Technically the definition of shift-invariance (which is also called difference scale invariance [8]) expresses stability of aggregation functions with respect to translation rather than invariance. Because the term shift-invariance is much in use, e.g. [7,9], we adopt it for the remainder of this paper.

2.2. Means

The term *mean* is used synonymously with averaging aggregation functions. Under the monotonicity constraint averaging behaviour and idempotency are equivalent, however without monotonicity, idempotency does not imply averaging. In this article we will use the term mean for averaging functions which are not necessarily monotone.