Expert Systems with Applications 38 (2011) 4908–4919

Contents lists available at ScienceDirect

Expert Systems with Applications

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

The centroid or consensus of a set of objects with qualitative attributes

Adolfo Guzman-Arenas *, Alma-Delia Cuevas, Adriana Jimenez

Centro de Investigación en Computación (CIC) and Escuela Superior de Cómputo (ESC), Instituto Politécnico Nacional (IPN), Mexico

ARTICLE INFO

Keywords:
I.2.4 Knowledge representation
Qualitative values
Inconsistency
Confusion
Consensus
Truth discovery

ABSTRACT

It is clear how to compute the average of a set of numeric values; thus, handling inconsistent measurements is possible. Recently, using *confusion*, we showed a new way to compute the consensus (a kind of average) of a set of assertions about a non-numeric fact, such as the religion of John.

This paper solves the same problem for a set of *objects* possessing several symbolic attributes. Suppose there is a murder, and we ask several observers about the height, sex, hair color and ethnicity of the killer. They report divergent observations. What is the most likely portrayal of the assassin? Given a bag of assertions about an object described by qualitative features, this paper tells how to assess the most plausible or "consensus" object description. It is the most likely description to be true, given the available information. It is the "centroid" of the bag. We also compute the *inconsistency* of the bag: how far apart the testimonies in the bag are. All observers are equally credible, so differences arise from perception errors, and from the limited accuracy of the individual findings.

1. Previous work

When measurements on the same quantitative attribute disagree, we resort to the average (or centroid) and variance of the results. We know how to take into account contradicting facts like these, and we do not regard them necessarily as inconsistent. We just assume that the measurers' gauges have different precisions or accuracies. It could also be that observers have a propensity to lie, and in this case we apply the Theory of Evidence (Dempster, 1968; Shafer, 1979). Or we could use Fuzzy Logic, selecting some sets as possible answers and assigning a degree of membership to each measurement for each set.

Yin, Han, and Fu (2008) provide a manner to find the most likely "truth" among a set of qualitative information[1] obtained from "information providers" in the Web. The information is an assertion about a qualitative value, a "fact" as found in the Web. This work resorts to the "trustworthiness" of each informant (resembling Dempster–Schafer), as well as a measure of the similarity among two of these non-numeric values (resembling our *confusion*, as defined in next pages).

A recent paper (Guzman-Arenas & Jimenez, 2010) finds the centroid or most likely value of a bag of qualitative values, such as {Afghanistan; Beirut; Iraq; Kabul; Middle East; Afghanistan; Syria}. The answer is not necessarily the most popular value or mode (Afghanistan), nor the least common ancestor (Middle East). The answer is not based on the probability that informants lie (like in the theory of evidence), nor it contains fuzzy values. The answer assumes that all informants are equally credible, and the discrepancy of their findings arises from the way or method used when obtaining their observations.

As an example, let us assume that we want to discover what pet Bart has, so we ask several observers to find out. One of them hears the animal bark, so he reports "a dog;" another observer finds fur hairs, so she reports "a mammal," while a third observer reports "a large dog," seeing the silhouette of the animal at night. Assume the reported values are {dog; mammal; German Shepherd; iguana}. One of them is the most likely pet. If we select "dog," reporter 1 is happy (he shows no discomfort, since our selection agrees with his report, a dog); reporter 2 is also happy (our selection agrees with her finding, a mammal); reporter 3 is somewhat displeased, since he observed a more accurate dog (a German Shepherd), not just "some dog." Reporter 4 is more uncomfortable, since he found an iguana. If our selection is "iguana", only reporter 4 is at comfort, while three others are somewhat upset. If we could measure these discomforts, we could select as the most likely pet (consensus value) *the value that minimizes the sum of disagreements* for all the observers when they learn of the value chosen as the consensus value.

The discomfort or disagreement when value *r* is reported instead of the "true" value *s* (as found by the observer) is called the *confusion* in using *r* instead of *s* (Levachkine, Guzman-Arenas, & de Gyves, 2005; Levachkine & Guzman-Arenas, 2007). To measure this, it is necessary to give all observers the same *context*, that is, the same set of possible qualitative answers as well as how these are related

* Corresponding author. Tel.: +52 55 5595 5075; fax: +52 55 5668 1250.
E-mail addresses: a.guzman@acm.org (A. Guzman-Arenas), almadeliacuevas@gmail.com (A.-D. Cuevas), dyidyia@yahoo.com (A. Jimenez).

[1] Qualitative attributes (such as *religion* or *hair color*) are also called non-numeric properties, aspects, features, or linguistic variables. The values these attribute attain (such as *Muslim* or *brown*) are called qualitative values, non numeric values, or linguistic constants.

0957-4174/$ - see front matter © 2010 Elsevier Ltd. All rights reserved.
doi:10.1016/j.eswa.2010.09.169

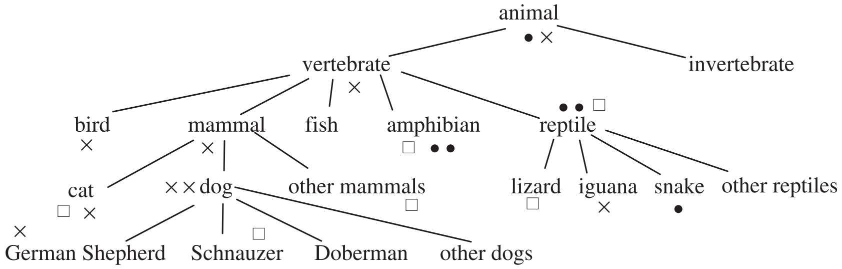

Fig. 1. A hierarchy of symbolic values is a tree where every node is either a symbolic value or, if it is a set, then its immediate descendants form a partition. Hierarchies make possible to compute the confusion conf (r,s) that results when value r is used instead of s, the true or intended value. The confusion (Section 1.1) is the number of *descending* links in the path from r to s, divided by the height of the hierarchy. For instance, conf (dog, Doberman) = 1/4, conf (Doberman, dog) = 0, conf (Doberman, German Shepherd) = 1/4, conf (Doberman, iguana) = 2/4, conf (iguana, Doberman) = 3/4. Observe that conf $\in [0,1]$. Refer to Section 1.1. The values marked ×, □ and ● are used in examples 5 and 6 of Section 1.2.

by specificity or generality. This set is called a *hierarchy* (Fig. 1); it is a tree where each node is a qualitative value or, if it is a set, then its immediate descendants form a *partition* of it.

Using hierarchies, next Section 1.1 tells how to compute the confusion among two qualitative values, while Section 1.2 explains how to find the consensus or most likely value of a bag of qualitative values.

Sections 1.1 and 1.2 report some previous work, necessary to understand this article. Our contributions appear in Section 2.1, which describes how to find the confusion when using an object O instead of the real or intended object O', and in Section 2.2, which obtains the consensus, centroid or most plausible object in a bag of objects, as well as the inconsistency of such bag, a number that reveals how disparate are its members.

1.1. Measuring the confusion between two qualitative values

Work on *confusion*has been reported elsewhere (Guzman & Levachkine, 2004a, 2004b; Levachkine & Guzman-Arenas, 2007; Levachkine et al., 2005); this section is placed here for completeness, in order to understand Section 2. How close are two numeric values v_1 and v_2? The answer is $|v_2–v_1|$. How close are two symbolic values such as *cat* and *dog*? The answer comes in a variety of similarity measures and distance functions. Hierarchies (introduced in Fig. 1) allow us to define the confusion conf (r,s) between two symbolic values. The function conf will open the way to evaluate in Section 1.2 the inconsistency of a bag of symbolic observations. We assume that the observers of a given fact (such as *the killer*) share a set of common vocabulary, best arranged in a hierarchy. A hierarchy can be regarded as the "common terminology"[2] for the observers of a bag: their *context*. Observers reporting in other bag may share a different context, that is, another hierarchy.

What is the capital of Germany? *Berlin* is the correct answer; *Frankfurt* is a close miss, *Madrid* a fair error, and *sausage* a gross error. What is closer to a *cat*, a *dog* or an *orange*? Can we measure these errors and similarities? Can we retrieve objects in a database that are close to a desired item? Yes, because qualitative variables take symbolic values such as *cat, orange, California, Africa*, which can be organized in a hierarchy H, a mathematical construct among these values. Over H, we can define the function *confusion*resulting when using a symbolic value instead of another.

Definition 1. For $r, s \in H$, the **absolute confusion** in using r instead of s, is

$$\mathrm{CONF}(r,r) = \mathrm{CONF}(r, \text{any ascendant of } r) = 0;$$
$$\mathrm{CONF}(r,s) = 1 + \mathrm{CONF}(r, \text{father\_of}(s)).$$

To measure CONF, count the descending links from r (the replacing value) to s (the intended or real value). CONF is neither a distance nor an ultradistance function.

We can normalize CONF by dividing it into h,the height of H (the number of links from the root of H to the farthest element of H), yielding the following.

Definition 2. The **confusion** in using r instead of s is conf (r,s) = CONF (r,s)/h.

Notice that $0 \leqslant \mathrm{conf}(r,s) \leqslant 1$. It is not symmetric: conf(r,s) $\neq$ conf(s,r), in general. The function conf is not a distance function, but it obeys the triangle inequality (Guzman-Arenas & Jimenez, 2010).

Example 1. For the hierarchy of Fig. 1, CONF (cat, mammal) = 0; if I ask for a mammal and I am given a cat instead, I am happy, and CONF = 0. But CONF (mammal, cat) = 1; if I ask for a cat and I get a mammal, I am somewhat unhappy, and CONF = 1. For the same reason, CONF (cat, vertebrate) = 2. Being given a vertebrate when I ask for a cat makes me unhappier than when I was handed a mammal.

Example 2. In the hierarchy of Fig. 1, conf (cat, dog) = 1/4; conf (cat, Schnauzer) = 1/2.

Remark 1. Since symbolic values lie in a hierarchy, it is not possible for a value to have two immediate ascendants, to have more than one path from it towards the root. That is, *rabbit* may not be both a mammal and a bird.

The type of hierarchy of Fig. 1 is the most common type, and it is sometimes called a *normal* hierarchy, as opposed to ordered (Section 1.1.1) or percentage hierarchies (Section 1.1.2).

1.1.1. When symbolic values are totally ordered

If the values in a hierarchy could be ordered by a " < " relation, for instance cold < chilly < tepid < warm < hot < burning, we will have an *ordered hierarchy*(Guzman & Levachkine, 2004b), with height 1 (Fig. 2) always. The confusion for ordered hierarchies with n children is:

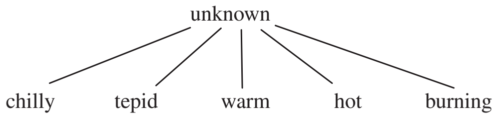

Fig. 2. An ordered hierarchy about temperatures, with root *unknown* and five children, which are ordered by a "<" relation. conf (*unknown*, *chilly*) = 1; conf (*warm*, *unknown*) = 0; conf (*chilly*, *tepid*) = conf (*tepid*, *chilly*) = conf (*tepid*, *warm*) = 1/4; conf (*chilly*, *warm*) = 1/2; conf (*hot*, *chilly*) = 3/4; conf (*chilly*, *burning*) = 1.

[2] If the symbolic values become full *concepts*, it is best to use an *ontology* instead of a *hierarchy* to place them (Cuevas et al., 2008).