Paper

Contextual probability

Hui Wang

Abstract — In this paper we present a new probability function G that generalizes the classical probability function. A *mass function* is an assignment of basic probability to some *context* (events, propositions). It represents the strength of support for some contexts in a domain. A *context* is a subset of the basic elements of interest in a domain – the *frame of discernment*. It is a medium to carry the "probabilistic" knowledge about a domain. The G function is defined in terms of a mass function under various contexts. G is shown to be a probability function satisfying the *axioms of probability*. Therefore G has all the properties attributed to a probability function. If the mass function is obtained from probability function by normalization, then G is shown to be a linear function of probability distribution and a linear function of probability. With this relationship we can estimate probability distribution from probabilistic knowledge carried in some contexts without any model assumption.

Keywords — mathematical foundations, knowledge representation, machine learning, uncertainty, data mining.

1. Introduction

Probability theory is the body of knowledge that enables us to reason formally about uncertain events or propositions. There are different approaches to probability theory, most notably the *frequentist* and *Bayesian* approaches [1, 4].

In the frequentist point of view, the probability of an event is taken to be equal to the limit of the relative frequency of the chosen event with respect to all possible events as the number of trials goes to infinity. The appeal of the frequentist approach for scientists lies in the apparent objectivity of its treatment of data.

On the other hand, the Bayesian approach extends the interpretation of probability to include degrees of belief or knowledge in propositions. We pass from the probability of events (frequentist) to the probability of propositions (Bayesian). Nevertheless the axioms used to define the mathematical properties of probability remain unchanged. Consequently many of the statistical procedures of the two approaches are identical.

Here we focus on the mathematical properties of probability. In particular we take probability to be defined in terms of probability distribution. Let Ω be a set consisting of the basic elements of interest in a domain. A probability distribution function is $p : \Omega \to [0,1]$ such that $\sum_{x\in\Omega} p(x) = 1$. A (classical) probability function is $P : 2^{\Omega} \to [0,1]$ such that, for any $E \subseteq \Omega$

$$P(E) = \sum_{x\in E} p(x). \tag{1}$$

The function $P(E)$ is the probability that an arbitrary element $x \in \Omega$ belongs to a well-defined subset $E \subseteq \Omega$.

It can be shown that the classical probability function defined above satisfies the *axioms of probability*: for any *event* $E \subseteq \Omega$:

- $P(E) \geq 0$.
- $P(\Omega) = 1$.
- If $E_1 \cap E_2 = \emptyset$ then $P(E_1 \cup E_2) = P(E_1) + P(E_2)$.

It is recognized that any function satisfying the axioms of probability, however defined, is a probability function [1].

If we know the probability distribution (in the case of finite sets) or density (in the case of infinite sets) we can calculate probability for any events – in a sense probability distribution provides us with complete information about a domain ([3], p. 273). Therefore probability distribution estimation – estimating the probability distribution from known probabilities for some events – is very important. This is in a sense a way of extending or generalizing our knowledge (represented by probabilities on some events) to all possible events of interest.

There are two general classes of distribution models: parametric and nonparametric. Parametric models assume a particular functional form for the distribution function, such as a uniform distribution, a normal distribution, and so on [3]. Parametric models are often characterized by a relatively small number parameters. Parametric models have the advantage of simplicity (easy to estimate and interpret) but may have relatively high bias because real data may not obey the assumed functional form.

In nonparametric models the distribution estimate is data-driven and relatively few assumptions are made *a priori* about the functional form. Histogram, kernel models and *k*-nearest-neighbors are examples. Histogram is a relatively primitive version of kernel method, and *k-NN* is a special case of kernel [5]. Kernel methods are based on the assumption that a function is constant locally, but the extent of "locality" is parameter to be given, which has a critical bearing on the performance of the methods.

In this paper we present a theory that generalizes the classical probability theory. It can estimate probability distribution without any model assumption.

2. Contextual probability function

Let Ω be a **finite** set called *frame of discernment*. $E \subseteq \Omega$ is called a *context* or *event*. A *mass function* is $m : 2^{\Omega} \to [0,1]$ such that

$$\sum_{X\subseteq\Omega} m(X) = 1. \tag{2}$$

The mass function is interpreted as a *representation of (probabilistic) knowledge* about Ω.

Our objective is to extend our knowledge to those contexts that we have no explicit knowledge about in m. Therefore we define a new function $G: 2^{\Omega} \to [0,1]$ such that for any $E \subseteq \Omega$

$$G(E) = \sum_{X \subseteq \Omega} m(X) \frac{|E \cap X|}{|X|}. \tag{3}$$

The interpretation of the above definition is as follows. Context E may not be known explicitly in the representation of our knowledge, but we know explicitly some contexts X that are related to it (i.e., E overlaps with X or $E \cap X \neq \emptyset$). Part of the knowledge about X ($m(X)$) should then be attributed to E. Since we do not know how this knowledge about X is distributed among the components in X, we can assume it is evenly distributed. So the part of this knowledge attributable to E is $m(X) \times |E \cap X|/|X|$.

Theorem 1. *G is a probability function on Ω. That is to say:*

1. *For any $E \subseteq \Omega$, $G(E) \geq 0$.*
2. *$G(\Omega) = 1$.*
3. *For $E_1, E_2 \in \Omega$, $G(E_1 \cup E_2) = G(E_1) + G(E_2)$ if $E_1 \cap E_2 = \emptyset$.*

Proof. The first claim is true following the fact that $m(X) \geq 0$ for any $X \subseteq \Omega$. The equation holds when $E = \emptyset$.

The second claim is true since $G(\Omega) = \sum_{X \subseteq \Omega} m(X)$.

Let's now consider the third claim. $X \cap (E_1 \cup E_2) = (X \cap E_1) \cup (X \cap E_2)$. If $E_1 \cap E_2 = \emptyset$ then $|X \cap (E_1 \cup E_2)| = |X \cap E_1| + |X \cap E_2|$. As a result we have

$$\begin{aligned} G(E_1 \cup E_2) &= \sum_{X \subseteq \Omega} m(X) \frac{|X \cap (E_1 \cup E_2)|}{|X|} = \\ &= \sum_{X \subseteq \Omega} m(X) \frac{|X \cap E_1| + |X \cap E_2|}{|X|} = \\ &= \sum_{X \subseteq \Omega} m(X) \frac{|X \cap E_1|}{|X|} + \sum_{X \subseteq \Omega} m(X) \frac{|X \cap E_2|}{|X|} = \\ &= G(E_1) + G(E_2). \end{aligned}$$

□

We therefore call G a *contextual probability function*, and the formalism about this function is termed *contextual probability theory* (CPT for short).

As a probability function G has the following properties, the proofs of which are left to the readers.

- $G(\bar{E}) = 1 - G(E)$.
- $G(\emptyset) = 0$.
- If $E_1 \subseteq E_2$, then $G(E_1) \leq G(E_2)$.
- $G(E) \leq 1$, for any context E.
- If $E_1, E_2, \cdots, E_n$ are contexts such that $E_i \cap E_j = \emptyset$ for all pairs i, j, then

$$G(\bigcup_{i=1}^{n} E_i) = \sum_{i=1}^{n} G(E_i).$$

- $G(E_1 \cup E_2) = G(E_1) + G(E_2) - G(E_1 \cap E_2)$.

By the above theorem we have

$$1 = G(\Omega) = \sum_{x \in \Omega} G(x).$$

As a result, if G is restricted to the singleton sets (i.e., the elements in Ω) it is a (calculated) probability distribution. This is in contrast to p, which can be interpreted as *a priori* probability distribution.

For simplicity, if E is a singleton set, e.g., $E = \{a\}$, we write $G(a)$ for $G(\{a\})$.

Now we look at an example before we move on.

Example 1. Let $\Omega = \{a,b,c,d,e,f\}$, and the mass function m be as follows:

$$\begin{aligned} m(\{a,b\}) &= 0.3 \\ m(\{a,b,c\}) &= 0.4 \\ m(\{a,b,c,d\}) &= 0.1 \\ m(\{a,b,c,d,e,f\}) &= 0.2 \end{aligned}$$

Suppose that we are interested in the probabilities of the contexts: $\{a\}, \{b\}, \{c\}, \{d\}, \{e\}, \{f\}, \{b,c\}, \{a,b,d\}$. According to the definition of G function, we have

$$\begin{aligned} G(a) &= m(\{a,b\}) \times \frac{|\{a\}|}{|\{a,b\}|} + m(\{a,b,c\}) \times \frac{|\{a\}|}{|\{a,b,c\}|} + \\ &\quad + m(\{a,b,c,d\}) \times \frac{|\{a\}|}{|\{a,b,c,d\}|} + \\ &\quad + m(\{a,b,c,d,e,f\}) \times \frac{|\{a\}|}{|\{a,b,c,d,e,f\}|} = \\ &= 0.3 \times 1/2 + 0.4 \times 1/3 + 0.1 \times 1/4 + 0.2 \times 1/6 = \\ &= 41/120 \end{aligned}$$

$$G(b) = G(a)$$

$$\begin{aligned} G(c) &= m(\{a,b,c\}) \times \frac{|\{c\}|}{|\{a,b,c\}|} + \\ &\quad + m(\{a,b,c,d\}) \times \frac{|\{c\}|}{|\{a,b,c,d\}|} + \\ &\quad + m(\{a,b,c,d,e,f\}) \times \frac{|\{c\}|}{|\{a,b,c,d,e,f\}|} = \\ &= 0.4 \times 1/3 + 0.1 \times 1/4 + 0.2 \times 1/6 = \\ &= 23/120 \end{aligned}$$

$$\begin{aligned} G(d) &= m(\{a,b,c,d\}) \times \frac{|\{d\}|}{|\{a,b,c,d\}|} + \\ &\quad + m(\{a,b,c,d,e,f\}) \times \frac{|\{d\}|}{|\{a,b,c,d,e,f\}|} = \\ &= 0.1 \times 1/4 + 0.2 \times 1/6 = 7/120 \end{aligned}$$

$$G(e) = m(\{a,b,c,d,e,f\}) \times \frac{|\{e\}|}{|\{a,b,c,d,e,f\}|} =$$
$$= 0.2 \times 1/6 = 4/120$$
$$G(f) = G(e).$$

Clearly $G(a)+G(b)+G(c)+G(d)+G(e)+G(f)=1$. Further on, we have

$$G(\{b,c\}) = m(\{a,b\}) \times \frac{|\{b\}|}{|\{a,b\}|} +$$
$$+ m(\{a,b,c\}) \times \frac{|\{b,c\}|}{|\{a,b,c\}|} +$$
$$+ m(\{a,b,c,d\}) \times \frac{|\{b,c\}|}{|\{a,b,c,d\}|} +$$
$$+ m(\{a,b,c,d,e,f\}) \times \frac{|\{b,c\}|}{|\{a,b,c,d,e,f\}|} =$$
$$= 0.3 \times 1/2 + 0.4 \times 2/3 + 0.1 \times 2/4 + 0.2 \times 2/6 =$$
$$= 64/120 = G(b) + G(c)$$
$$G(\{a,b,d\}) = m(\{a,b\}) \times \frac{|\{a,b\}|}{|\{a,b\}|} +$$
$$+ m(\{a,b,c\}) \times \frac{|\{a,b\}|}{|\{a,b,c\}|} +$$
$$+ m(\{a,b,c,d\}) \times \frac{|\{a,b,d\}|}{|\{a,b,c,d\}|} +$$
$$+ m(\{a,b,c,d,e,f\}) \times \frac{|\{a,b,d\}|}{|\{a,b,c,d,e,f\}|} =$$
$$= 0.3 + 0.4 \times 2/3 + 0.1 \times 3/4 + 0.2 \times 3/6 =$$
$$= 89/120 = G(a) + G(b) + G(d).$$

3. CPT versus probability theory and Dempster-Shafer theory

Contextual probability theory generalizes classical probability theory in the sense that the probability distribution p changes to the mass function m and the probability function P changes to contextual probability function G. The probability distribution p is defined on Ω while m is defined on 2^Ω; P and G are both defined on 2^Ω and they are both probability functions. Once the mass function is restricted to singletons the G function becomes the probability function.

Dempster-Shafer (D-S for short) theory [6] is also a generalization of probability theory, which has evolved from a theory of upper and lower probabilities. It starts by assuming a set Ω and a *mass function* m, based on which the *belief* function *bel* and *plausibility* function *pls* are defined. Formally the mass function is $m : 2^\Omega \to [0,1]$ where $m(\emptyset) = 0$ and $\sum_{X \subseteq \Omega} m(X) = 1$. Belief function is $bel : 2^\Omega \to [0,1]$ such that, for $E \subseteq \Omega$, $bel(E) = \sum_{X \subseteq \Omega, X \subseteq E} m(X)$. Plausibility function is $pls : 2^\Omega \to [0,1]$ such that $pls(E) = 1 - bel(E')$, where E' is the complement of E in Ω.

The set Ω is a set of mutually exclusive alternatives. For any $E \subseteq \Omega$, $m(E)$ represents the strength of some evidence supporting E; $bel(E)$ summarizes all reasons to believe E, and $pls(E)$ expresses how much we should believe in E if all currently unknown facts were to support E. Thus the true belief in E will be somewhere in the interval $[bel(E), pls(E)]$.

When the mass function is restricted to singleton elements $x \in \Omega$, the belief and plausibility functions become the same and they are also the same as the probability function. Therefore D-S theory is regarded as a generalization of probability theory [2].

However the belief function satisfies the first two axioms of probability theory, but for the third axiom the equation is changed to "$\geq$" [2]. Therefore the belief function is not probability function.

Although both CPT and D-S theory can be understood as generalizations of probability theory, there are differences between the two:

- CPT uses a single function to represent "uncertainty" while D-S theory uses two functions.
- The G function is a probability function, therefore all of the properties of probability theory are still valid. For example, with the additive property we do not need to calculate G for every $E \subseteq \Omega$; instead we only need to do so for singletons $x \in \Omega$ and $G(E) = \sum_{x \in E} G(x)$. The belief function is, however, not a probability function. So we have to calculate *bel* for every $E \subseteq \Omega$.

4. Relationship between G and P

Now that G is a probability function, we may ask the question: what is the relationship between G and P? To answer this question we need to base m on p so that G can be connected to P. Here we interpret mass function as a measure of the occurrence of elements in a set. Therefore the larger a set is the more likely the set as an event occurs.

Specifically we assume that the mass function be defined in terms of probability as follows, letting $K \stackrel{\text{def}}{=} \sum_{X \subseteq \Omega} P(X)$

$$m(E) = \frac{P(E)}{\sum_{X \subseteq \Omega} P(X)} = \frac{P(E)}{K}. \tag{4}$$

According to this interpretation of the mass function the following lemma follows from the fact that $P(E_1) \leq P(E_2)$ when $E_1 \subseteq E_2$.

Lemma 1. *If* $E_1 \subseteq E_2$ *then* $m(E_1) \leq m(E_2)$.

Let $\binom{N}{n}$ be the combinatorial number representing the number of ways of picking n unordered outcomes from N possibilities. From combinatorics we know that $\binom{N}{n} = \frac{N!}{(N-n)!n!}$. With these assumptions we have the following results.

Lemma 2. *Let* $N = |\Omega|$. *Then* $K = \sum_{i=1}^{N} \binom{N-1}{i-1} = 2^{N-1}$.

Proof.

$$K = \sum_{X\subseteq\Omega} P(X) = \sum_{i=1}^{N} \sum_{X\subseteq\Omega,|X|=i} P(X) =$$
$$= \sum_{i=1}^{N} \sum_{x\in\Omega} \binom{N-1}{i-1} p(x) = \sum_{i=1}^{N} \binom{N-1}{i-1} \sum_{x\in\Omega} p(x) =$$
$$= \sum_{i=1}^{N} \binom{N-1}{i-1} = 2^{N-1}.$$

□

Theorem 2. *Let* $\alpha \stackrel{\text{def}}{=} \frac{1}{K}\sum_{i=1}^{N} \frac{\binom{N-2}{i-1}}{i}$, *and* $\beta \stackrel{\text{def}}{=} \frac{1}{K}\sum_{i=1}^{N} \frac{\binom{N-2}{i-2}}{i}$. *Then* $G(x) = \alpha p(x) + \beta$ *for* $x \in \Omega$.

Proof.

$$G(x) = \sum_{Y\subseteq\Omega} \frac{x\cap Y}{|Y|} m(Y) = \sum_{Y\subseteq\Omega, x\in Y} \frac{1}{|Y|}\frac{P(Y)}{K} =$$
$$= \frac{1}{K} \sum_{Y\subseteq\Omega, x\in Y} \frac{P(Y)}{|Y|} = \frac{1}{K} \sum_{Y\subseteq\Omega, x\in Y} \frac{\sum_{z\in Y} p(z)}{|Y|} =$$
$$= \frac{1}{K} \sum_{i=1}^{N} \frac{1}{i} \sum_{Y\subseteq\Omega,|Y|=i, x\in Y} \sum_{z\in Y} p(z) =$$
$$= \frac{1}{K} \sum_{i=1}^{N} \frac{1}{i} \left(\binom{N-1}{i-1} p(x) + \binom{N-2}{i-2} \sum_{z\neq x} p(z) \right) =$$
$$= \frac{1}{K} \sum_{i=1}^{N} \frac{1}{i} \left(\binom{N-1}{i-1} p(x) + \binom{N-2}{i-2} (1-p(x)) \right) =$$
$$= \frac{1}{K} \sum_{i=1}^{N} \frac{1}{i} \left(\binom{N-2}{i-1} p(x) + \binom{N-2}{i-2} \right) =$$
$$= \alpha p(x) + \beta.$$

□

The claim then follows.

Since both P and G are probability functions we have $\sum_{x\in\Omega} P(x) = 1$ and $\sum_{x\in\Omega} G(x) = 1$. According to Theorem 2 we then have:

Corollary 1. $\alpha + |\Omega| \times \beta = 1$.

As a result we only need to calculate either of α and β, and the other can be determined according to the corollary.

Since both P and G are probability functions they satisfy the additive axiom. In other words for $E \subseteq \Omega$, $P(E) = \sum_{x\in E} p(x)$ and $G(E) = \sum_{x\in E} G(x)$. Following Theorem 2 we then have:

Corollary 2. $G(E) = \alpha P(E) + \beta |E|$.

Theorem 2 and Corollary 2 establish the relationship between G and probability distribution, and G and probability respectively. If we have full knowledge about the distribution we can calculate probability, which can further be used to calculate G. On the contrary, if we have full knowledge about G then we can calculate distribution and probability *precisely*. The interesting question is, if we have only incomplete or partial knowledge about G then we can get an approximation to the probability (and probability distribution). Therefore CPT can be used as a method for probability distribution estimation.

Example 2. Consider a set $\{a,b,c,d\}$, whose probability distribution is assumed to be $\{0.1, 0.3, 0.4, 0.2\}$. Following definition, the P, m and G values can be calculated for all the subsets (contexts) and are shown in Table 1.

Table 1
The set is $\Omega = \{a,b,c,d\}$. The probability distribution is assumed to be $\Omega = \{a:0.1, b:0.3, c:0.4, d:0.2\}$. Note that $p(x) = P(\{x\})$ for $x \in \Omega$

E	∅	{a}	{b}	{c}
P	0	0.1	0.3	0.4
m	0	1/80	3/80	4/80
G	0	198/960	254/960	282/960
E	{d}	{a,b}	{a,c}	{a,d}
P	0.2	0.4	0.5	0.3
m	2/80	4/80	5/80	3/80
G	226/960	452/960	480/960	424/960
E	{b,c}	{b,d}	{c,d}	{a,b,c}
P	0.7	0.5	0.6	0.8
m	7/80	5/80	6/80	8/80
G	536/960	480/960	508/960	734/960
E	{a,b,d}	{a,c,d}	{b,c,d}	{a,b,c,d}
P	0.6	0.7	0.9	1.0
m	6/80	7/80	9/80	10/80
G	678/960	706/960	762/960	1

Clearly the G values for singleton subsets are slightly different from the those P values given in probability distribution.

Let's now illustrate the relationship between P and G with respect to Theorem 2.

Here Ω has four elements so $N = 4$. Then according to Lemma 2, $K = 2^{N-1} = 2^3 = 8$. Other components in the theorem are calculated as follows:

$$\alpha = \frac{1}{K}\sum_{i=1}^{N} \frac{\binom{N-2}{i-1}}{i} =$$
$$= \frac{1}{8}\left(\binom{2}{0} + \frac{\binom{2}{1}}{2} + \frac{\binom{2}{2}}{3} + \frac{\binom{2}{3}}{4} \right) = \frac{1}{8}\left(1 + 1 + \frac{1}{3}\right) = \frac{28}{96}$$

$$\beta = \frac{1}{K}\sum_{i=1}^{N} \frac{\binom{N-2}{i-2}}{i} =$$
$$= \frac{1}{8}\left(\binom{2}{-1} + \frac{\binom{2}{0}}{2} + \frac{\binom{2}{1}}{3} + \frac{\binom{2}{2}}{4} \right) = \frac{1}{8}\left(\frac{1}{2} + \frac{2}{3} + \frac{1}{4}\right) = \frac{17}{96}.$$

Now we show that $P(x) = [G(x) - \beta]/\alpha$ for every elements in the set.

$$[G(a) - \beta]/\alpha = \left(\frac{198}{960} - \frac{17}{96}\right) \times \frac{96}{28} = \frac{28}{960} \times \frac{96}{28} = 0.1$$

$$[G(b) - \beta]/\alpha = \left(\frac{254}{960} - \frac{17}{96}\right) \times \frac{96}{28} = \frac{84}{960} \times \frac{96}{28} = 0.3$$

$$[G(c) - \beta]/\alpha = \left(\frac{282}{960} - \frac{17}{96}\right) \times \frac{96}{28} = \frac{112}{960} \times \frac{96}{28} = 0.4$$

$$[G(d) - \beta]/\alpha = \left(\frac{226}{960} - \frac{17}{96}\right) \times \frac{96}{28} = \frac{56}{960} \times \frac{96}{28} = 0.2$$

Clearly the equation holds.
To illustrate Corollary 2, consider context $E = \{a,b,c\}$. By the additive property of the G function we have

$$G(E) = G(a) + G(b) + G(c) = 734/960.$$

By the definition of G we have

$$\begin{aligned} G_1(E) &= m(a) + m(b) + m(c) + m(\{a,b\}) + m(\{a,c\}) + \\ &+ m(\{b,c\}) + m(\{a,b,c\}) + \frac{3}{4} \times m(\{a,b,c,d\}) = \\ &= \frac{1}{80} + \frac{3}{80} + \frac{4}{80} + \frac{4}{80} + \frac{5}{80} + \frac{7}{80} + \frac{8}{80} + \frac{3 \times 10}{4 \times 80} = \\ &= \frac{158}{320} \end{aligned}$$

$$\begin{aligned} G_2(E) &= \frac{1}{2} m(\{a,d\}) + \frac{1}{2} m(\{b,d\}) + \frac{1}{2} m(\{c,d\}) = \\ &= \frac{3}{2 \times 80} + \frac{5}{2 \times 80} + \frac{6}{2 \times 80} = \frac{14}{160} \end{aligned}$$

$$\begin{aligned} G_3(E) &= \frac{2}{3} m(\{a,b,d\}) + \frac{2}{3} m(\{a,c,d\}) + \frac{2}{3} m(\{b,c,d\}) = \\ &= \frac{2 \times 6}{3 \times 80} + \frac{2 \times 7}{3 \times 80} + \frac{2 \times 9}{3 \times 80} = \frac{44}{240} \end{aligned}$$

$$G(E) = G_1(E) + G_2(E) + G_3(E) = \frac{734}{960}.$$

The probability function $P(E)$ is calculated according the additive property as follows:

$$P(E) = P(\{a\}) + P(\{b\}) + P(\{c\}) = 0.8.$$

Using the α and β values above we have

$$P(E) \times \alpha + |E| \times \beta = 0.8 \times \frac{28}{96} + 3 \times \frac{17}{96} = \frac{734}{960} = G(E)$$

Example 3. Now let's look at another example with the same Ω and the same underlying probability distribution. We assume that we do not know the distribution explicitly, but we know the probability values for some subsets. These values are normalized to give a mass function. The known probability and mass values are shown in Table 2.

Table 2
The probability and mass values for Example 3

E	{a,b}	{a,c}	{b,d}	{c,d}
P	0.4	0.5	0.5	0.6
m	4/50	5/50	5/50	6/50
E	{a,b,c}	{a,b,d}	{a,c,d}	{b,c,d}
P	0.8	0.6	0.7	0.9
m	8/50	6/50	7/50	9/50

By definition we have

$$\begin{aligned} G(a) &= \frac{1}{2}[m(\{a,b\}) + m(\{a,c\})] + \\ &+ \frac{1}{3}[m(\{a,b,c\}) + m(\{a,b,d\}) + m(\{a,c,d\})] = \\ &= \frac{1}{2} \times \frac{4+5}{50} + \frac{1}{3} \times \frac{8+6+7}{50} = \frac{69}{300} = 0.230 \end{aligned}$$

$$\begin{aligned} G(b) &= \frac{1}{2}[m(\{a,b\}) + m(\{b,d\})] + \\ &+ \frac{1}{3}[m(\{a,b,c\}) + m(\{a,b,d\}) + m(\{b,c,d\})] = \\ &= \frac{1}{2} \times \frac{4+5}{50} + \frac{1}{3} \times \frac{8+6+9}{50} = \frac{73}{300} = 0.243 \end{aligned}$$

$$\begin{aligned} G(c) &= \frac{1}{2}[m(\{a,c\}) + m(\{c,d\})] + \\ &+ \frac{1}{3}[m(\{a,b,c\}) + m(\{a,c,d\}) + m(\{b,c,d\})] = \\ &= \frac{1}{2} \times \frac{5+6}{50} + \frac{1}{3} \times \frac{8+7+9}{50} = \frac{81}{300} = 0.270 \end{aligned}$$

$$\begin{aligned} G(d) &= \frac{1}{2}[m(\{b,d\}) + m(\{c,d\})] + \\ &+ \frac{1}{3}[m(\{a,b,d\}) + m(\{a,c,d\}) + m(\{b,c,d\})] = \\ &= \frac{1}{2} \times \frac{5+6}{50} + \frac{1}{3} \times \frac{6+7+9}{50} = \frac{77}{300} = 0.257 \end{aligned}$$

To calculate P values from respective G values we need K, α and β, which are functions of N. From Example 2 we know that $K = 8$, $\alpha = 28/96 = 0.292$ and $\beta = 17/96 = = 0.177$. $P(x)$ can be calculated by $P(x) = [G(x) - \beta]/\alpha$ for every elements as follows:

$$P(\{a\}) = (0.230 - 0.177)/0.292 = 0.18$$

$$P(\{b\}) = (0.243 - 0.177)/0.292 = 0.23$$

$$P(\{c\}) = (0.270 - 0.177)/0.292 = 0.32$$

$$P(\{d\}) = (0.257 - 0.177)/0.292 = 0.27$$

With these values we can calculate probability for any other subsets.

If we apply Dempster-Shafer theory we can calculate the belief and plausibility values for any contexts. For example

$$bel(\{x\}) = 0 \text{ for any } x \in \Omega.$$

The plausibility value is calculated as follows:

$$pls(\{a\}) = m(\{a,b\}) + m(\{a,c\}) + m(\{a,b,c\}) + \\ + m(\{a,b,d\}) + m(\{a,c,d\}) = 30/50$$

$$pls(\{b\}) = m(\{a,b\}) + m(\{b,d\}) + m(\{a,b,c\}) + \\ + m(\{a,b,d\}) + m(\{b,c,d\}) = 32/50$$

$$pls(\{c\}) = m(\{a,c\}) + m(\{c,d\}) + m(\{a,b,c\}) + \\ + m(\{a,c,d\}) + m(\{b,c,d\}) = 35/50$$

$$pls(\{d\}) = m(\{b,d\}) + m(\{c,d\}) + m(\{a,b,d\}) + \\ + m(\{a,c,d\}) + m(\{b,c,d\}) = 33/50.$$

5. Summary and conclusion

In this paper we have presented a new probability function G – contextual probability function, which is defined in terms of a basic probability assignment – mass function. Therefore G has all the properties of the classical probability function, which satisfies the three *axioms of probability*. The mass function has similar meaning as that in the D-S theory. Thus CPT enjoys the flexibility and other properties attributed to the D-S theory. The key difference between the two, however, is the fact that G is a probability function whereas the belief and plausibility functions are not. One consequence is that, due to the additive property, only the G values for singleton elements in Ω need to be calculated and the G values for any other subsets of Ω can be obtained from the G values for singletons. This is a big save in time. Since the belief and plausibility functions in the D-S theory are not additive we have to calculate belief and plausibility values for all subsets of Ω.

Though G is a probability function, CPT can be viewed as a generalization of the classical probability theory in the sense that if the mass function is defined only for elements in Ω, the G becomes the P function. The D-S theory is also regarded as a generalization of probability theory, but the belief and plausibility functions are not probability functions; they become probability functions when the mass function is defined only for elements in Ω.

The mass function can be interpreted in different ways for different purposes. We have shown that if it is interpreted as normalized (summing up to 1) probability function, G is a linear function of the P function. This connection makes it possible to estimate probability distribution from the probability values of some known events.

Future work should include interpreting the mass function in other ways for other purposes and applying the CPT to some real world problems.

References

[1] W. Feller, *An Introduction to Probability Theory and its Applications*. Wiley, 1968.

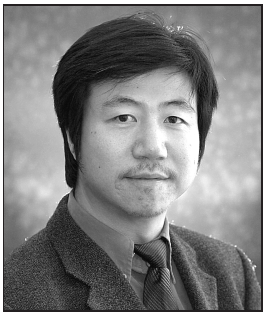
[2] J. W. Guan and D. A. Bell, "Generalization of the Dempster-Shafer theory" in *Proc. IJCAI-93*, 1993, pp. 592–597.

[3] D. Hand, H. Mannila, and P. Smyth, *Principles of Data Mining*. The MIT Press, 2001.

[4] E. T. Jaynes, "Probability theory: the logic of science", http://bayes.wustl.edu

[5] B. D. Ripley, *Pattern Recognition and Neural Networks*. Cambridge University Press, 1996.

[6] G. Shafer, *A Mathematical Theory of Evidence*. Princeton, New Jersey: Princeton University Press, 1976.



Hui Wang is a senior lecturer at the University of Ulster, Northern Ireland, UK. He received his B.Sc. and M.Sc. from Jilin University (China) and Ph.D. from the University of Ulster. His research interests include machine learning, data mining, text classification, probability theory, qualitative reasoning, neural networks, and Internet technologies.

e-mail: H.Wang@ulster.ac.uk
School of Computing and Mathematics
Faculty of Informatics
University of Ulster
Newtownabbey, Northern Ireland
BT 37 0QB, United Kingdom