Document Type Recognition Using Evidence Theory

Thierry Géraud[1], Geoffroy Fouquier[1], Quôc Peyrot[1],

Nicolas Lucas[2], Franck Signorile[2]

[1] EPITA Research and Development Laboratory, France

`www.lrde.epita.fr`

[2] SWT, France

`www.swt.fr`

Who's who?

EPITA Research and Development Laboratory:

- software engineering,
- scientific computing in C++, meta-programming
- image processing, pattern recognition.

SWT:

- French company, editor of the "b-Wize" software product line
  "solutions to sort, index, read, retrieve and process contents from paper sources"
- winner of the European IST Prize 2003

`http://www.ist-prize.org/`

Outline

- introduction —context and intentions
- a running example
- first solutions:
  - Boolean logic approach
  - fuzzy approach
- evidence theory:
  - basics
  - modeling
  - comparative results
- conclusion and perspectives

Context

Document type recognition:

- document types are known —a type database/knowledge base exists
- type = set of characteristics
- a characteristic can be featured by several document types
- evaluation "characteristic c / document d" $\Rightarrow$ value $\in [0, 1]$
  - 1 means "d *does* feature c"
  - 0 means "d *does not* feature c"
  - 0.5 means "d *more or less* features c"

Example of characteristics:

- a flower-shaped logo is on top-left (W)
- document font is 12pt (F)
- there is a bar code somewhere (B)
- etc.

Intents

Within *this* context, we do *not* explain:

- how to build such a knowledge base
- how to select relevant features
- how to valuate couples such as “a characteristic / a document”.

We focus on how to handle information to proceed to document type recognition.

Keywords:

- information management
  - fusion
  - imprecision
- decision
  - uncertainty
  - conflict

evidence theory is not new but is not well-known $\rightarrow$ let us be didactic...

Running example

characteristics	document types			documents		
	type 1 (t_1)	type 2 (t_2)	type 3 (t_3)	case 0 (d_0)	case 1 (d_1)	case 2 (d_2)
flower logo (W)	yes	no	no	no	0.1	0.2
12pt fonts ($F_{1}{}^{2}$)	yes	no	yes	no	0.8	0.7
bar code (B)	no	yes	yes	no	0.7	0.5

This example is simple enough to be quickly solvable by a human.

Real applications are far more complicated:

- many characteristics,
- many document types,
- most of the characteristics are featured by several document types

Boolean logic 1/2

	type 1 (t_1)	type 2 (t_2)	type 3 (t_3)	case 0 (d_0)
flower logo (W)	true	false	false	false
12pt fonts (F_{1^2})	true	false	true	false
bar code (B)	false	true	true	false

Notation:

- $1_{t_i}(d)$ = "d has type t_i"
- $1_{t_i}(c_j)$ = "c_j is a characteristic of t_i"
- $1_{c_j}(d)$ = "d features c_j"

$$1_{t_i}(d) = \bigwedge_j \left(1_{t_i}(c_j) = 1_{c_j}(d) \right)$$

Example:

$1_{t_2}(d_0)$ is false since d_0 and t_2 does not perfectly match.

Boolean logic 2/2

Main drawbacks:

- decisions are taken too early
- errors are propagated

No proper way to:

- handle imprecision
- measure ambiguity

Definitions:

- *Imprecision*: lack of precise knowledge (syn. inaccuracy).
- *Uncertainty*: incomplete knowledge.
- *Vagueness*: lack of clearness in contours or limits.
- *Fusion*: mixing several pieces of information.

Fuzzy approaches:

- well suited to model these notions
- decision is taken at the very end.

Fuzzy set theory

D: set of documents
$S_i \subset D$: *fuzzy* subset of D
$d \in D$: a document

$\mu_{S_i}(d) \in [0,1]$: membership degree
$\cup_i S_i = D \quad \Rightarrow \quad \sum_i \mu_{S_i}(d) = 1$
(normalization)

Fuzzy sets derived from characteristics:

$$\left\{ \begin{array}{lcl} W & = & W_{yes} \cup W_{no} \\ F & = & F_{12} \cup \overline{F_{12}} \\ B & = & B_{yes} \cup B_{no} \end{array} \right. \Rightarrow \left\{ \begin{array}{l} \text{scheme 1:} \\ D = W \times F \times B \quad \Rightarrow \quad t_1 = W_{yes} \times F_{12} \times B_{no} \\ \text{or} \\ \text{scheme 2:} \\ D = W = F = B \quad \Rightarrow \quad t_1 = W_{yes} \cap F_{12} \cap B_{no} \end{array} \right.$$

where $\overline{F_{12}} = F_{<12} \cup F_{>12}$

denoting c_i^j the subset of (characteristic) c^j corresponding to t_i :

when either $\left(t_i = \bigtimes_j c_i^j \right)$ or $\left(t_i = \bigcap_j c_i^j \right)$, we have: $\underline{\mu_{t_i}(d) = \min_j \mu_{c_i^j}(d)}$.

Fuzzy fusion

Generalization with a fuzzy fusion operator:

$$\mu_{t_i}(d) = \bigoplus_j \mu_{c_i^j}(d)$$

$\oplus$ can be conjunctive:

- "deciding to assign d to t_i means that we *simultaneously* well recognize every features c_i^j in document d"
- Conjunctive operators are T-norms and verify $\oplus \leq \min$.

$\oplus$ can be a compromise:

- "deciding to assign d to t_i means that we *globally* well recognize all features c_i^j in the document d"
- Compromise operators are means and verify $\min < \oplus < \max$ (between T-norms and T-conorms).

Fuzzy decision

Decision function:

$$\underline{\omega(d) = \arg\max_{i} \mu_{t_i}(d)}$$

2nd best decision: $\omega_2(d) = \arg \max_{i \neq \omega(d)} \mu_{t_i}(d)$

No decision is taken when:

- confidence is too low, i.e. $\mu_{t_{\omega(d)}} < h_1$
- ambiguity is noticed, i.e.
  - $\mu_{t_{\omega(d)}} - \mu_{t_{\omega_2(d)}} < h_2$

    or
  - $\frac{\mu_{t_{\omega(d)}}}{\mu_{t_{\omega_2(d)}}} < h_3.$

Fuzzy fusion results

	t_1	t_2	t_3	$\mu_{*\text{yes}}(d_1)$	$\mu_{*\text{yes}}(d_2)$
W	yes	no	no	0.1 $\rightarrow$ no	0.2 $\rightarrow$ no
F_{12}	yes	no	yes	0.8 $\rightarrow$ yes	0.7 $\rightarrow$ yes
B	no	yes	yes	0.7 $\rightarrow$ yes	0.5 $\rightarrow$?
			intuitive results	$\rightarrow t_3$	$\rightarrow t_3$

d_1

	t_1	t_2	t_3
min	0.10	0.20	0.70
mean	0.40	0.60	0.80
μ	0.22	0.33	0.44

d_2

	t_1	t_2	t_3
min	0.20	0.30	0.50
mean	0.47	0.53	0.67
μ	0.28	0.32	0.40

where μ is the normalized arithmetical mean.

Temporary conclusion *1/2*

When $\oplus$ is conjunctive,
false estimations of feature presence can lead to false results;

$\oplus$ should be a compromise but then
a lot of false ambiguities appear...

Temporary conclusion 2/2

Main problem:

- different types can have several characteristics in common;
- until now, each document type is handled separately;
- actually we valuate singletons...

A simple illustration:

- set of people = $\{$ Greg, Jack, Tom $\}$
- statement = “ I can’t remember who’s the biggest fool but I’m positive that it’s either Greg or Tom. ”
  - Fuzzy modeling = $0.5_{/\text{Greg}} + 0.5_{/\text{Tom}} + 0_{/\text{Jack}}$
  - Drawback = 0.5 for Greg means “half a fool”
  - Proper translation = $1_{/(\text{Greg or Tom})} + 0_{/\text{Jack}}.$

Evidence theory 1/3

Hypothesis set: $\Theta = \{\, t_1, \, \ldots, \, t_n \,\}$.

Mass function:

$$\begin{cases} \forall A \subset \Theta, \quad m(A) \in [0,1] \\ \sum_{A \subset \Theta} m(A) = 1 \\ m(\emptyset) = 0. \end{cases}$$

$A \subset \Theta$ is a focal element if $m(A) \neq 0$.

Several functions $A \subset \Theta \to [0,1]$ are defined.

Belief function (amount of evidence which implies A):

$$bel(A) = \sum_{B \subset A} m(B).$$

Uncertainty about A:

interval $[\, bel(A), \, pls(A) \,]$

Ignorance: $ign(A) = pls(A) - bel(A)$.

Plausibility function (amount of evidence that does not refute A):

$$pls(A) = 1 - bel(\overline{A}) = \sum_{B \cap A \neq \emptyset} m(B).$$

Doubt about A (amount of evidence that does refute A):

$$dou(A) = bel(\overline{A}).$$

Evidence theory 2/3

Measure of conflict between s sources ($m_i,\ i = 1..s$):

$$K = \sum_{\cap_{i=1}^{s} B_i = \emptyset} \left(\prod_{i=1}^{s} m_i(B_i) \right).$$

Mass combination (Dempsters's rule):

$$\left(\bigoplus_{i=1}^{s} m_i \right)(A) \ = \ \frac{1}{1-K} \sum_{\cap_{i=1}^{s} B_i = A} \left(\prod_{i=1}^{s} m_i(B_i) \right).$$

Property:

$$m = \bigoplus_{i=1}^{s} m_i \ \text{ is a mass.}$$

Finally, we compute from m:

$$\forall i, \quad bel(\{t_i\}) \ \text{ and } \ pls(\{t_i\}).$$

Evidence theory 3/3

Decision rules

- maximum of belief:

$$\omega_{bel}(d) = \arg\max_i \; bel(\{t_i\})(d)$$

- maximum of plausibility:

$$\omega_{pls}(d) = \arg\max_i \; pls(\{t_i\})(d)$$

- absolute decision rule = maximum of belief without overlapping of belief intervals:

$$\omega_{abs}(d) = \omega_{bel}(d) \quad \underline{\text{if}} \quad \forall i \neq \omega_{bel}(d),\; pls(\{t_i\})(d) < bel(\{t_{\omega_{bel}(d)}\})(d)$$

- compromise = maximum of $(bel + pls)/2$:

$$\omega_{\mathit{cpm}}(d) = \arg\max_i \; \frac{bel + pls}{2}(\{t_i\})(d)$$

Evidence modeling

With global uncertainty:

	type 1 (t_1)	type 2 (t_2)	type 3 (t_3)	focal elements	
flower logo (W)	yes	no	no	$m_W(\{t_1\})$	$m_W(\Theta)$ $^{(*)}$
12pt fonts (F_{1^2})	yes	no	yes	$m_{F_{12}}(\{t_1, t_3\})$	$m_{F_{12}}(\Theta)$
bar code (B)	no	yes	yes	$m_B(\{t_2, t_3\})$	$m_B(\Theta)$

$^{(*)}$ this means: "according to W, when it is not t_1, it is either t_1, t_2, or t_3";
we then have: $m_W(\Theta) = 1 - m_W(\{t_1\})$.

Fusion step:

$$m_u = m_W \oplus m_{F_{12}} \oplus m_B .$$

Without global uncertainty:

e.g., $m_{\not{t}W}(\{t_1\}) = m_W(\{t_1\})$ and $m_{\not{t}W}(\Theta - \{t_1\}) = m_W(\Theta)$
means: "according to W, when it is not t_1, it is either t_2 or t_3".

Results 1/2

Three different approaches $\Rightarrow$ results having three different flavors.

d_1

	$\{t_1\}$	$\{t_2\}$	$\{t_3\}$	$\{t_1, t_3\}$	$\{t_2, t_3\}$	$\{t_1, t_2, t_3\}$
m_u	0.03	0.00	0.54	0.23	0.14	0.06
$m_{\not u}$	0.04	0.19	0.77	0.00	0.00	0.00
μ	0.22	0.33	0.44	undef	undef	undef

d_2

	$\{t_1\}$	$\{t_2\}$	$\{t_3\}$	$\{t_1, t_3\}$	$\{t_2, t_3\}$	$\{t_1, t_2, t_3\}$
m_u	0.11	0.00	0.31	0.31	0.13	0.13
$m_{\not u}$	0.15	0.26	0.60	0.00	0.00	0.00
μ	0.28	0.32	0.40	undef	undef	undef

Results 2/2

Comparison "fuzzy / evidence" (decision = compromise)

d_1

	$\{t_1\}$	$\{t_2\}$	$\{t_3\}$
bel	0.03	0.00	0.54
pls	0.32	0.19	0.97
evidence	0.18	0.10	0.75
fuzzy	0.22	0.33	0.44

d_2

	$\{t_1\}$	$\{t_2\}$	$\{t_3\}$
bel	0.11	0.00	0.31
pls	0.56	0.27	0.89
evidence	0.33	0.13	0.60
fuzzy	0.28	0.32	0.40

Conclusion

Evidence theory:

- is well-suited to handle both imprecision and uncertainty in document type recognition;
- allows to describe document types by (fuzzy) characteristics.

Effective application:

- several thousand documents to be processed;
- about one hundred different document types;
- quasi-perfect recognition results.

Implementation

Materials:

- we provide free software libraries under the GNU PUBLIC LICENSE (GPL)
- downloadable from `www.lrde.epita.fr`

Mathematical Theory of Evidence

- project `eVidenZ`

Image Processing and Pattern Recognition

- project `Olena`

Thanks for your attention; any questions?