Извлечение информации из текста с автоматическим построением правил

Прокофьев П.А., Васильев В.Г.

20 октября 2011 г.

Извлечение информации из текстов

- **Извлечение информации из текстов** (ИИ) связано с получением описанных в тексте объектов, их атрибутов и взаимных связей
- [Carreras,2002], [Tsukamoto,2002], **[Curran,2003]**, [Krishnan,2006], [Smith,2006], **[Turmo,2006]**, [Funayama,2009], [Андреев,2007], [Куршев,2007], [Котельников,2010], [Алексеев,2009]
- **Классификация фрагментов текстов** — ключевой момент ИИ
- Эксперименты на задаче разрешения одного вида неоднозначностей при географической привязке текстов [Прокофьев, 2009]

Цель исследования

- **Объектом** исследования являются тексты и фрагменты текстов
- **Предметом** исследования являются методы и инструменты извлечения информации (ИИ)
- **Цель** — формализовать задачи, возникающие на различных этапах ИИ, и разработать гибко настраиваемые экспертами инструменты, использующие математические методы обработки информации

Задачи исследования

Задачи:

1) исследовать используемые в задачах ИИ признаки фрагментов;
2) исследовать, адаптировать или разработать методы поиска хороших закономерностей для алгоритмов классификации;
3) обеспечить описание закономерностей доступным экспертам языком правил;
4) автоматизировать построение правил;
5) оценить качество предложенных методов на прикладной задаче

Признаки фрагментов в задачах ИИ

[Carreras,2002], [Tsukamoto,2002], [Malouf,2002], [Collins,2002], **[Curran,2003]**, [Krishnan,2006], [Smith,2006], [Sasano,2008], [Funayama,2009], [Бородин,2008], [Алексеев,2009]:

- признаки сравнения слов;
- признаки морфологических, графематических, синтаксических, семантических дескрипторов;
- признаки по словарям;
- признаки локального контекста;
- признаки глобального контекста;
- признаки выполнения регулярных выражений Regex;
- признаки выполнения сложных правил;

Модель текстов и правил

- Текст $\tau = (\tau_1, \ldots, \tau_L) \in \mathbb{T}$, τ_i — контекст слова
- Фрагмент текста

$$S = \tau[i,j], |S| = j - i + 1, \mathbb{F}(\tau) = \{\tau[i,j] | 1 \leq i \leq j \leq L\}$$

- Правило

$$q(\tau) = \{\tau[i_1, j_1], \ldots, \tau[i_c, j_c]\}$$

- Примеры правил

$$q^*(\tau) = \mathbb{F}(\tau); q^{[n,m]}(\tau) = \{S \in \mathbb{F}(\tau) \mid n \leq |S| \leq m\}$$

- Использование в правилах предикатов с параметрами

$$q^{(g,\alpha)}(\tau) = \{\tau[i,j] | g(\alpha, (\tau_i, \ldots, \tau_j)) = 1\}, g : A \times \mathbb{T} \to \{0,1\}$$

Множественные операторы правил

- Операторы пересечения, объединения, дополнения и разности:

$$(q_1 \between q_2)(\tau) = q_1(\tau) \cap q_2(\tau);$$

$$(q_1 \bigtriangledown q_2)(\tau) = q_1(\tau) \cup q_2(\tau);$$

$$(\neg q)(\tau) = \mathbb{F}(\tau) \setminus q(\tau);$$

$$(q_1 \ominus q_2)(\tau) = q_1(\tau) \setminus q_2(\tau);$$

- `В | НА` $\longrightarrow$
  `КОНФЕРЕНЦИЯ [В] ВОРОНЕЖЕ НАЗНАЧЕНА [НА] ОКТЯБРЬ ...;`
- [Агеев М.С.,2005]

Оператор логического «И»

- Склейка фрагментов:

$$\tau[i_1, j_1] \triangle \tau[i_2, j_2] = \tau[\min\{i_1, i_2\}, \max\{j_1, j_2\}];$$

- Оператор И:

$$(q_1 \triangle q_2)(\tau) = \{S_1 \triangle S_2 | S_1 \in q_1(\tau), S_2 \in q_2(\tau)\}$$

- В & РАЙОН $\longrightarrow$
  СОВЕЩАНИЕ [В ДМИТРОВСКОМ РАЙОНЕ] ...;
  ДМИТРОВСКИЙ [РАЙОН СНИЗИЛ В] МАЕ ...

Операторы префиксного и суффиксного условий

- Упорядочим фрагменты:

$$\tau_1[i_1, j_1] <_{n,m} \tau_2[i_2, j_2] \equiv n \leq i_2 - j_1 \leq m,$$

- Префиксное правило:

$$(\square^{\leftarrow}_{n,m} q)(\tau) = \{S | \exists S_1 \in q(\tau), S_1 <_{n,m} S\}$$

- Суффиксное правило:

$$(\square^{\rightarrow}_{n,m} q)(\tau) = \{S | \exists S_1 \in q(\tau), S <_{n,m} S_1\}$$

- `В ?:1:2 $СУЩ` $\longrightarrow$
  `МЕРОПРИЯТИЕ В [РАЙОНЕ] [ГОРОДА] ...`

  `ВЛАДИМИР :! ЛЕНИН` $\longrightarrow$
  `В ГОРОДЕ [ВЛАДИМИРЕ] ОТКРЫТ НОВЫЙ МУЗЕЙ.`

Оператор ограничения пространства

- Ограничение фрагмента:

$$\tau[i_1, j_1] \sqsupset \tau[i_2, j_2] \equiv i_1 \leq i_2 \leq j_2 \leq j_1$$

$$\forall A \subset \mathbb{F}(\tau), \triangleleft_{i,j}(A) = \{S \in A | S \sqsubset \tau[i, j]\};$$

- Ограничение правила:

$$(\triangleleft_{i,j} q)(\tau) = \begin{cases} \triangleleft_{i,j}((*_1(\triangleleft_{i,j} q_1))(\tau)), q = *_1 q_1, \\ \triangleleft_{i,j}(((\triangleleft_{i,j} q_1) *_2 (\triangleleft_{i,j} q_2))(\tau)), q = q_1 *_2 q_2, \\ \triangleleft_{i,j}(q(\tau)), \text{иначе}, \end{cases}$$

где $*_1$ и $*_2$ — унарная и бинарная операции;

- Ограничение правила по правилу:

$$(q_1 \triangleleft q_2)(\tau) = \bigcup\nolimits_{\tau[i,j] \in q_2(\tau)} (\triangleleft_{i,j} q_1)(\tau)$$

- `В : ГОРОД` $\longrightarrow$
  `НАЧАЛО [В ОКТЯБРЕ. ГОРОД] ВОРОНЕЖ ...`
  `В :\s ГОРОД` $\longrightarrow$
  `НАЧАЛО В ОКТЯБРЕ. ГОРОД ВОРОНЕЖ ...`

Алгебраические свойства правил

- Элементарные свойства:

$$q_1 \between (q_2 \bigtriangledown q_3) = (q_1 \bigtriangledown q_2) \between (q_1 \bigtriangledown q_3);$$

- Специфические свойства:

$$(\Box^{\leftarrow}_{n_1,m_1} q) \between (\Box^{\leftarrow}_{n_2,m_2} q) = \Box^{\leftarrow}_{\max\{n_1,n_2\},\min\{m_1,m_2\}} q;$$

- Ослабление или усиление правила:

$$\Box^{\leftarrow}_{n,1} q(\tau) \supset \Box^{\leftarrow}_{n,2} q(\tau)$$

Классификация фрагментов, описанных правилами

- Фрагменты являются классифицируемыми объектами:

$$\bigcup_{\tau \in \mathbb{T}} \mathbb{F}(\tau) = \mathbb{K}_1 \sqcup \ldots \sqcup \mathbb{K}_m$$

- Обучение по выборке размеченных текстов:

$$X = K_1 \sqcup \ldots \sqcup K_m, K_i = \mathbb{K}_i \cap X, i = 1, \ldots, m$$

- Классификатор работает с векторами значений признаков:

$$(f_1(S), \ldots, f_n(S))$$

- Признаки задаются правилами (например, [Curran,2003]):

$$f^{(q)}(\tau[i,j]) = \begin{cases} 1, \text{если } \tau[i,j] \in q(\tau), \\ 0, \text{иначе.} \end{cases}$$

- Правилами языка описаны признаки из [Carreras,2002], [Tsukamoto,2002], [Curran,2003]

Логические (дискретные) процедуры распознавания

- [Журавлев Ю.И.,1966,1981], [Вайнцвайг М.Н.,1971], [Дюкова Е.В.,2005], [Песков Н.В.,2004]
- Процедура состоит из наборов элементарных классификаторов для классов и функции голосования
- Представительные наборы голосуют ЗА:

$$\Gamma_1^{A_1}(\vec{\beta}, K_i) = \sum\nolimits_{c \in C^{A_1}(K_i)} \gamma(c) B(c, \vec{\beta}).$$

- Антипредставительные — НАОБОРОТ:

$$\Gamma_2^{A_2}(\vec{\beta}, K_i) = \sum\nolimits_{c \in C^{A_2}(K_i)} \gamma(c)(1 - B(c, \vec{\beta})).$$

- Вместе голосуют ЛУЧШЕ:

$$\Gamma_3^{A_3}(\vec{\beta}, K_i) = \Gamma_1^{A_1}(\vec{\beta}, K_i) + \Gamma_2^{A_2}(\vec{\beta}, K_i)$$

Обучение процедур распознавания

- Элементарный классификатор на языке логических функций:

$$x_{i_1}^{\alpha_1} \cdot \ldots \cdot x_{i_r}^{\alpha_r}$$

- Обучение поиском сокр. ДНФ частично заданной логической функции [С.Б. Яблонский,1974]. Например, для представительных наборов:

$$u^{(K_i,X)}(\vec{\beta}) = \begin{cases} 1, \vec{\beta} \in K_i \\ 0, \vec{\beta} \in X \setminus K_i. \end{cases}$$

- Поиск (p, q)- представительных наборов [Н.В. Песков,2004]. Как выбрать p и q?

Преимущества и недостатки дискретных процедур

(+) Элементарный классификатор описывается правилом и может стать признаков:

$$x_1 \longrightarrow q_1; x_2 \longrightarrow \square^{\leftarrow}_{n,n} q_2; x_3 \longrightarrow \square^{\rightarrow}_{m,m} q_3,$$

$$\text{тогда } x_1^1 x_2^1 x_3^0 \longrightarrow q_2 \, \square^{\leftarrow}_{n,n} \, q_1 \, \square^{\nrightarrow}_{m,m} \, g_3$$

(—) Критичность к числу признаков и прецедентов: не более 70 признаков на текущей задаче

(—) Необходимость выбирать поднабор признаков: генетические алгоритмы, **Add-Del**

(+) Обучение может быть продолжено итеративно

Эксперимент на прикладной задаче

Задача разрешения неоднозначностей типов географических объектов:

- 846 текстов, содержащих 372367 слов;
- 5431 фрагментов отмечены одним из 8 типов ;
- 6408 «подозрительных» фрагментов, то есть 977 — «другие», итого 9 классов;
- 56 «базовых» правил, отобранных Add-Del;
- оценка методов $MaxEnt(ME)$, Γ_1, Γ_2, $\Gamma_1 + ME$, $\Gamma_2 + ME$, $\Gamma_1 + \Gamma_2$
- оценка кросс-валидацией: 3 разбиениями со стратификацией по классам, усреднение точности и полноты;
- 3 итерации усложнения правил с оценкой после каждой итерации

Результаты экспериментов

Таблица: Оценка качества методов

Метод	P	R	F	e
ME	42,5%	37,6%	35,2%	17,4%
Γ_1	76,2%	65,8%	63,3%	15,6%
Γ_2	38,3%	24,8%	21,7%	27,2%
$\Gamma_1 + ME$	76,1%	59,8%	54,9%	22,2%
$\Gamma_2 + ME$	30,0%	21,3%	19,9%	25,5%
$\Gamma_1 + \Gamma_2$	**77,1%**	**72,6%**	**69,6%**	**11,2%**

Таблица: Оценка качества для хорошего класса «Область»

Метод	P	R	F	e
Γ_1	94,3%	98,0%	96,2%	0,9%
$\Gamma_1 + \Gamma_2$	94,3%	98,0%	96,2%	0,3%

Построенные правила

Таблица: «Базовые» правила

Правило	Информативность
@ОБЛАСТЬ	0,98
#1 :1? РАЙОН	0,36
#1 :1? КРАЙ	0,82
ЕДИНАЯ РОССИЯ	0,05

Таблица: «Составные» правила

Текст правила	Классификатор
@ОБЛАСТЬ :1! РАЙОН	$x_1^1 x_2^0$
@ОБЛАСТЬ :1! {$Verb}	$x_1^1 x_3^0$

Логические процедуры для трансформации признаков

Выводы, позволившие улучшить результат:

- представительные и антипредставительные наборы являются хорошими зависимостями;
- строить логические процедуры по небольшому набору информативных (**IGain**) признаков;
- использовать для классификации топовые алгоритмы (**SVM**, **C4.5** и др.);
- использовать бустинг для итеративного улучшения признаков-правил

Эксперимент на новом подходе:

- выборка та же;
- 5 итераций бустинга с выбором 30 лучших признаков для построения логической процедуры;
- оценка методов NB, SVM, $C4.5$, $C4.5 + Boosting$, $OneR$, KNN, $SVM + \Gamma_1$, $C4.5 + \Gamma_1$, $SVM + \Gamma_1 + \Gamma_2$, $C4.5 + \Gamma_1 + \Gamma_2$

Результаты экспериментов с новым подходом

Таблица: Оценка качества методов

Метод	P	R	e
$\Gamma_1 + \Gamma_2$	**77,1%**	**72,6%**	**11,2%**
NB	80,53%	95,35%	8,39%
OneR	82,64%	72,66%	11,32%
KNN	92,38%	86,42%	5,58%
*C*4.5	92,22%	90,99%	4,08%
*C*4.5 + *Boosting*	92,22%	90,99%	4,08%
*C*4.5 + Γ_1	92,89%	91,18%	3,94%
*C*4.5 + $\Gamma_1 + \Gamma_2$	89,05%	87,65%	3,93%
SVM	91,56%	86,88%	3,87%
SVM + Γ_1	93,91%	92,87%	3,26%
SVM + $\Gamma_1 + \Gamma_2$	**94,66%**	**92,48%**	**3,16%**

Выводы и заключение

- использование логических процедур распознавания не дало хороших результатов на стоящей прикладной задаче;
- построение логических процедур может сопровождаться поиском хороших зависимостей, описание которых возможно на предложенном языке правил;
- набор признаков может быть трансформирован в набор хороших зависимостей;
- деревья решений или веса SVM в совокупности с зависимостями, описанными на понятном экспертам языке, могут хорошо интерпретироваться экспертами;
- при этом правила могут быть настроены экспертами;

Дальнейшие исследования:

- расширение языка правил и исследование алгебраических свойств;
- описание правилами других признаков, использующихся в работах посвященных ИИ