О ёмкости семейств характеристических функций, обеспечивающих корректное решение задач диагностического типа

М.И.Забежайло

ФИЦ ИУ РАН

Интерполяционно-экстраполяционная схема

- *ИЭ*-схема для формализации задач машинного обучения (**неклассическая *интерполяция* + *экстраполяция*** по Ю.И.Журавлеву) :

Даны:

- множество Ω^+ (примеров) *объектов*, обладающих целевыми *свойствами* $\boldsymbol{P}_i^+$ из некоторого множества $\boldsymbol{P} = \{ P_{i1}, P_{i2}, \dots, P_{ir} \}$, и
- множество Ω^- (контрпримеров) объектов, не обладающих целевыми свойствами из $\boldsymbol{P} = \{ P_{i1}, P_{i2}, \dots, P_{ir} \}$,
- *новый объект* O^τ (или же некоторое явным образом представленное множество таких объектов Ω^τ).

Требуется:

оценить *наличие* (или отсутствие) *целевых свойств* у нового *объекта* O^τ (новых объектов из заданного множества Ω^τ), т.е.

- дать соответствующий ***прогноз*** (о наличии целевых свойств) и
- предъявить ***основания*** (**неоспоримые *аргументы***), позволяющие ***принять*** этот прогноз

- **Пример**: задачи диагностического типа

Эффект переобучения

Эффект переобучения как критичная характеристика *ИЭ*-схем (*К.В.Воронцов, Д.В.Виноградов*)

- **Воронцов К.В**. Комбинаторная теория надёжности обучения по прецедентам. – Дисертация на соискание учёной степени д.ф-м.н. по специальности 05.13.17 Теоретические основы информатики. – М: ВЦ РАН. - 273 С. – URL: https://www.dissercat.com/content/kombinatornaya-teoriya-nadezhnosti-obucheniya-po-pretsedentam

- **Виноградов Д.В.** Вероятностно-комбинаторный формальный метод обучения, основанный на теории решеток. – Дисертация на соискание учёной степени д.ф-м.н. по специальности 05.13.17 Теоретические основы информатики. – М.:ФИЦ ИУ РАН, 2018. – 131 с. – URL: http://www.frccsc.ru/diss-council/00207305/diss/list/vinogradov_dv

Можно ли бороться с переобучением?

Как бороться с переобучением? В чем его причина?
Возможные версии ответа:

- ***Каузальность*** vs ***интерполируемость***
- Таблица прецедентов и «объясняющие» ее частично определенные функции для задач диагностического типа
- *Проблема* ***устойчивости*** интерполирующих ЭЗ при расширении БФ описаниями новых прецедентов
  - «универсальные» ЭЗ
  - контекстно-зависимые ЭЗ

Опыт поиска «универсальных» ЭЗ

Опыт поиска «универсальных» эмпирических зависимостей ЭЗ в ИИ-исследованиях 1990-2000-х годов (**D.Lenat**, **Y.Zytkow**, **R.Michalski**, …)

- **Jan M.Zytkow** University of North Carolina, Dept. of Computer Science. Charlotte, NC 28223, USA

  Principles of data mining and knowledge discovery : Proc. 3-d European conference “PKDD ’99”, Prague, Czech Republic, September 15 - 18, 1999. ***Jan M. Zytkow*** ; ***Jan Rauch*** (ed.). - ***Lect. notes in comp. science.*** - Vol. 1704 : ***Lecture notes in artificial intelligence***. ISBN 3-540-66490-4. -Berlin et al.: Springer, 1999. - https://link.springer.com/content/pdf/bfm%3A978-3-540-48247-5%2F1.pdf

- **D.Lenat** - https://en.wikipedia.org/wiki/Douglas_Lenat

  Lenat D.B., Fishwick P.A., Modjeski R.B., Oresky C.M., Clarkson A., Kaisler S. (1991). "**STRADS: A Strategic Automatic Discovery System**" .*Knowledge-based Simulation: Methodology and Application*. – Pp. 133-161. - https://www.springer.com/gp/book/9780387973746 (См. Главу 11: https://www.springer.com/gp/book/9780387973746)

- **R.Michalski** - https://en.wikipedia.org/wiki/Ryszard_S._Michalski

  Michalski, R.S., Carbonell, J.G., Mitchell, T.M.: ‘**Machine learning: an artificial intelligence approach**’ (Springer Science & Business Media, Berlin, Germany, 2013). - https://www.springer.com/gp/book/9783662124079

Проблема устойчивости ЭЗ при расширении анализируемой Базы Фактов

- ***Big Data***: эффекты ***Big*** и ***Open***
- Последовательности *расширяющихся* Баз Фактов
- Проблема *устойчивости* ЭЗ, «интерполирующих» данную БФ, в условиях характерного для *Big Data* эффекта *Open* – *открытости\пополняемости* текущей БФ новыми эмпирическими фактами (**В.К.Финн**)

Особые свойства задач диагностического типа

- Необходимость оперировать ***выборками ограниченного размера*** (по количеству прецедентов), что ставит под вопрос ***адекватность*** применения целого ряда методов ***статистического анализа данных***
- Возможное отсутствие оснований рассматривать наблюдаемые (используемые для описания прецедентов) параметры как ***независимые***
- Возникновение эффектов ***Big Data*** на уровне размеров множеств значений фиксируемых параметров
- Проблемы с ***однородностью*** анализируемых прецедентов (и их описаний)
- ***Интерпретируемость*** в содержательных терминах и понятиях исследуемой предметной области (а не сформированной для ее описания алгоритмической модели) ***результатов*** (*заключений*, *диагнозов*, …), формируемых использованием тех или иных математических инструментов анализа данных
- Неформальная ***объясняемость*** результатов (*заключений*, *диагнозов*, …)
- ***Персонализация*** заключений (*диагнозов*): возможности учесть *индивидуальные особенности* пациента
- Проблема ***доверия*** - ***достаточности оснований*** для ***принятия*** порождаемых заключений (*диагнозов*, врачебных *рекомендаций* и т.п.)
- Наличие в рассматриваемой предметной области ***доказуемо трудно-разрешимых*** переборных задач, накладывающее существенные ограничения на **практическую применимость** ряда математических методов анализа данных и поддержки принятия диагностических решений

Характеристические Функции на последовательностях расширяющихся БФ в задачах диагностического типа

- Характеристические Функции как (*каузально-ориентированный*) подкласс семейства частичных функций, интерполирующих выборки прецедентов – примеров и контрпримеров диагностируемого явления

- Характеристическая Функция (**ХФ**), принимает
  - значение «*истина*» на всех фактах **ф** (примерах) текущей базы фактов **БФ**, характеризуемых наличием анализируемого целевого свойства и
  - значение «*ложь*» на всех фактах **ф** (контрпримерах) текущей базы фактов **БФ**, характеризуемых наличием анализируемого целевого свойства

$$\forall \mathbf{ф}\ [\ (\mathbf{ф} \in \mathbf{БФ}) \supset (\ \mathbf{ХФ}(\mathbf{БФ})\ |\!\!-\ \mathbf{ф})]$$

Процедура порождения Характеристических Функций

Прецеденты (формализованное описание)

=>

Сходство прецедентов (формализация)

=>

Классы сходства прецедентов

=>

Классы эквивалентности прецедентов

=>

Запрет на контрпримеры (условие ***ЗКП***)

=>

Покрытия текущей БФ **ЗКП-классами эквивалентности** примеров

- Пусть $\mathbf{\Omega} = \{ O_1, O_2, \dots , O_m \}$. Построим по $\mathbf{\Omega}$ и $\otimes$ множество $Dom(\mathbf{\Omega})$:

  (i) $\mathbf{\Omega} = \{O_1, O_2, \dots , O_m \} \subset Dom\ (\mathbf{\Omega})$.

  (ii) $\{[A \in Dom(\mathbf{\Omega})] \,\&\, [B \in Dom\ (\mathbf{\Omega})] \,\&\, [(A \otimes B) \neq \ulcorner \varnothing \urcorner]\} \rightarrow [\ (A \otimes B) \in Dom\ (\mathbf{\Omega})\ ]$,

  (iii) Других элементов в $Dom\ (\mathbf{\Omega})$ нет.

- Отношение сходства $\mathbf{R}^{\otimes}$:

$$O_{i1}\mathbf{R}^{\otimes}O_{i2}\ (\text{т.е.} < O_{i1}, O_{i2} > \in \mathbf{R}^{\otimes}) \quad \text{iff} \quad O_{i1} \otimes O_{i2} \neq \ulcorner \varnothing \urcorner$$

- Классы сходства (толерантности) на $\mathbf{\Omega}$:

$$\mathbf{T}(O_{i1}) = \{O_{i2} |\ O_{i1}\mathbf{R}^{\otimes}O_{i2}\}$$

- Фиксируя каждый конкретный результат $V=V_0$ вычисления операции $\otimes$ сходства на элементах множества $Dom\ (\mathbf{\Omega})$:

$$A \otimes B = V_0$$

  выделим соответствующие подклассы (эквивалентности) сформированных классов сходств:

$$\mathbf{E}_{V0} = \{< O_{i1}, O_{i2} > |\ O_{i1} \otimes O_{i2} \otimes V_0 = V_0\}$$

- Запрет на контрпримеры (условие ***ЗКП***):

$$(\forall V_0)\ [\ \big(\ (\mathbf{E}^{+}_{V0} = \{< O^{+}_{i1}, O^{+}_{i2} > |\ (O^{+}_{i1} \otimes O^{+}_{i2} \otimes V_0 = V_0)\}) \,\&\, (O^{+}_{i1} \in \mathbf{\Omega}^{+}) \,\&\, (O^{+}_{i2} \in \mathbf{\Omega}^{+})\ \big) \supset$$

$$\supset \neg(\exists\, O^{-}_{0})\ \{(V_0 \otimes O^{-}_{0} = V_0\ ) \,\&\, (O^{-}_{0} \in \mathbf{\Omega}^{-})\}\ ]$$

Процедура формирования Характеристических Функций

Определение

Будем называть характеристической функцией $\text{ХФ}_i(\mathbf{БФ})$ базы фактов **БФ** логическое условие, описывающее конкретное покрытие Cov_i (**БФ**) по следующей схеме:

- для каждого класса эквивалентности прецедентов $\text{GC-ЗКП}_{ij}(\mathbf{БФ})$, входящего в покрытие Cov_i (**БФ**) анализируемого множества прецедентов $\mathbf{БФ} = \boldsymbol{\Omega} = \boldsymbol{\Omega}^+ \cup \boldsymbol{\Omega}^-$, (при этом, естественно, $\boldsymbol{\Omega}^+ \cap \boldsymbol{\Omega}^- = \varnothing$), формирующее его множество признаков $\{a^{ij}_1, a^{ij}_2, \dots, a^{ij}_{n(ij)}\}$ - образующих из исходного множества $\boldsymbol{U} = \{a_1, a_2, \dots, a_n\}$ - порождает в сопоставляемой покрытию Cov_i (**БФ**) характеристической функции $\text{ХФ}_i(\mathbf{БФ})$ конъюнкцию переменных –

$$Q_{ij} = p^{ij}_1 \,\&\, p^{ij}_2 \,\&\dots\& \, p^{ij}_{n(ij)} \,, \qquad \text{a}$$

- сама характеристическая функция $\text{ХФ}_i(\mathbf{БФ})$ представляет собою дизъюнкцию по всем (сопоставляемым входящим в покрытие Cov_i (**БФ**) классам эквивалентности $\text{GC-ЗКП}_{ij}(\mathbf{БФ})$) конъюнкциям Q_{ij} :

$$\text{ХФ}_i(\mathbf{БФ}) = \bigvee_{\text{по всем j}} (Q_{ij}).$$

Характеристические Функции и каузальная репрезентативность анализируемой БФ

Утверждение 1 (о свойствах ХФ как характеристической функции, разделяющей примеры и контрпримеры с учетом причин возникновения анализируемого целевого эффекта)

Каждая построенная в соответствии с **Определением 1** функция $ХФ_i$(**БФ**) принимает на факте **ф** из текущей **БФ** значение

– «*истина*» тогда и только тогда, когда данный факт характеризуется наличием анализируемого целевого свойства;

- «*ложь*» тогда и только тогда, когда данный факт характеризуется отсутствием анализируемого целевого свойства,

и наоборот:

- на каждом факте **ф** из текущей **БФ**, характеризуемом наличием анализируемого целевого свойства, характеристическая функция $ХФ_i$(**БФ**) принимает значение «*истина*», а значение «ложь» тогда и только тогда, когда данный факт характеризуется отсутствием анализируемого целевого свойства

- **Непустота множества** ХФ(**БФ**) как характеристика *каузальной репрезентативности* анализируемой БФ

Задача о емкости семейства всех ХФ, порождаемых на данной БФ

Задача о емкости семейства ***ХФ***(**БФ**) всех ХФ, порождаемых на данной **БФ**:

Дана: база фактов (множество прецедентов) $\mathbf{БФ} = \boldsymbol{\Omega} = \boldsymbol{\Omega}^{+} \cup \boldsymbol{\Omega}^{-}$,
объединяющая примеры $\boldsymbol{\Omega}^{+}$ и контрпримеры $\boldsymbol{\Omega}^{-}$

Требуется: определить, сколько элементов содержит множество ***ХФ***(**БФ**)
всех ХФ, порождаемых на данной **БФ**

Утверждение 2

Задача о числе характеристических функций, формируемых на произвольной базе фактов **БФ**, принадлежит классу ***#PC*** перечислительно полных переборных задач

□

Задача о емкости семейства минимальных (по вложимости покрытий) ХФ

Задача о числе ХФ, *минимальных* по вложимости соответствующих им покрытий БФ *ЗКП*-классами эквивалентности примеров из БФ:

Дана: база фактов (множество прецедентов) $\mathbf{БФ} = \boldsymbol{\Omega} = \boldsymbol{\Omega}^{+} \cup \boldsymbol{\Omega}^{-}$,
объединяющая примеры $\boldsymbol{\Omega}^{+}$ и контрпримеры $\boldsymbol{\Omega}^{-}$

Требуется: определить, сколько элементов содержит множество $\boldsymbol{ХФ}_{min}(\mathbf{БФ})$ всех ХФ, порождаемых минимальными по вложению покрытиями данной **БФ** *ЗКП*-классами эквивалентности примеров из $\boldsymbol{\Omega}^{+}$

Утверждение 3

Задача о числе характеристических функций, формируемых на произвольной базе фактов *минимальными* по вложению покрытиями данной **БФ** *ЗКП*-классами эквивалентности примеров принадлежит классу $\#\boldsymbol{PC}$ перечислительно полных переборных задач

□

«Быстрая» разрешимость задач о ХФ?

Проблема *быстроты диагностирования*:

- каузальной репрезентативности анализируемой **БФ**
- непустоты подсемейства ХФ, наследуемых (*сохраняющих способность корректно представлять и* **БФ** *и ее конкретное расширение* **БФ** + Δ**БФ**) при добавлении к текущей **БФ** новых прецедентов Δ**БФ**

Заключение

- **Каузальность** $\neq$ **интерполируемость**

  (особый статус задач диагностического типа)

- Комбинаторная «**плата**» за точность интерполяции в задачах диагностического типа

Спасибо за внимание

m.zabezhailo@yandex.ru