Mining patterns on tabular data

François Amat

May 20, 2021

Overview

Candidate

François Amat (`https://famat.me`)

Graduate of Télécom Paris (2019), currently employed at Dassault Systèmes as a Data scientist. I am passionate about symbolic AI and knowledge bases.

Academia

- Graduate of the M2 Data&Knowledge, Saclay (2019)
- Graduate of the Engineering degree Télécom Paris (2019)
- Several research internships in France and abroad.

Industrial

- 2 years as data scientist at Dassault Systèmes
- Built products with knowledge extraction from wikidata
- Constructing a joint thesis proposal with Fabian Suchanek

Goal

Understand tabular data

Find patterns that are interesting to humans

Input: Car accidents from NHTSA (Open data)

Car model	Year	Death
Pathfinder	1994	0
Pontiac	1993	1
Lexus ES250	1993	0

Desired output:

- Deaths are **NOT** linked with the **Car model**.
- If the car is 5 years old, the death rate increases by 10%.
- Deaths are linked with the part **Seat bealt:front:anchorage**.

Issue

Deep learning or other black box models cannot deliver.

Example: Legal compliance

Use case

Let's suppose that I am the head of legal.

Input:

Company and open data

Topic	Legal code	Risk
DOL	CIVIL	Low
DOL	INSURANCE	Medium
DOL	FISCAL	High

Table: Open tabular data from https://www.legifrance.gouv.fr/

Desired output:

Insights such as :

- Arguing **DOL** is very risky for **Fiscal** issues.
- From 2010 to 2020 arguing **DOL** in **CIVIL** has increase failure by **21%**.

Example: Drug reimbursement policy

Use case

Let's suppose that I am working at the FRENCH social welfare.

Input:

open data

Name	Progress	%
PRALUENT	no	65
FUCIDINE	N/A	0
VERZENIOS	yes	100

Table: Open tabular data from `https://www.has-sante.fr/`

Desired output:

Check if there is evidence for patterns of interest such as :

- Company name → high reimbursement
- lack of progress → high reimbursement

Example: Human Resources

Use case

Let's suppose that I am the head of Human Ressources.

Input:

Name	Gender	Salary
Greg	Male	$50,078
Michael	Male	$276,500
Karen	Female	$240,000

Table: Open tabular data from `https://www.salaries.texastribune.org/`

Desired output:

Check if there is evidence for patterns of interest such as :

- If candidate age > 50 Then final acceptance ratio is $< 10\%$.
- If candidate ethnicity is minority Then final acceptance ratio is LOWER than other candidates.

Related work: Explainable AI (xAI) - Interpretable models

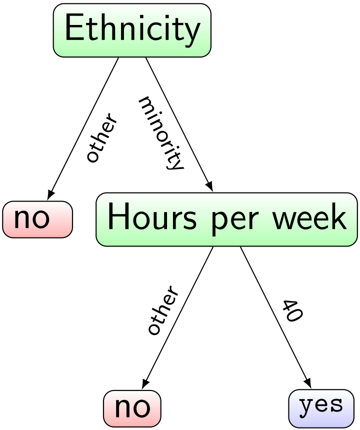


Figure: Decision tree

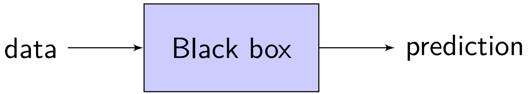
Classical Interpretable models are decision trees [8], rule-based models [13] and linear models [12].

Limitations :

They cannot find relations across multiples rows.

Is the data compliant with our HR policy, in that the director of an employee is always a manager?

Related work : Explainable AI (xAI) - Post-Hoc Models



Post-Hoc models aim to find limits, outlines of the prediction, or to map a black-box model to an understandable model.

Limitations :

- When they map to an understandable model they have the same limitations as these understable models (previous slides).
- They cannot explain how the outline of one prediction is made.

Related work: Inductive Logic Programming (ILP)

ILP is the task of learning logical rules from positive and negative examples. ILP methods find logical rules of the form :

IF $\mathit{relation}_1$(X,Y) and $\mathit{relation}_2$(Y,Z) THEN $\mathit{relation}_3$(X,Z)

Limitations of ILP

- Does not scale to millions of facts
- Has trouble dealing with negations under the open world assumption.

Related work: Rule mining

Rule mining is ILP designed to scale to millions of facts in large knowledge bases, under the open world assumption.

Limitations of Rule mining

- Cannot find numerical correlations.
- Cannot use predicates with arity > 2.
- Cannot collect rules with existential quantifiers.

Capabilities	XAI	Rule mining
Scalable to millions of entities	False	True
Explanable	True	True
Work under open world assumption	False	True
Combine several data points	False	True
Handle Negations	False	To improve
Work with tabular data	True	False
Work with arbitrary pattern	False	True
Existential quantifiers	False	False

Figure: State of the art

Rule mining system, AMIE as a basis

Association rule **M**ining under **I**ncomplete **E**vidence (**AMIE**).

AMIE has been developed at Telecom Paris since 2013.
AMIE is open source [1] and aims to be the reference and leader in rule mining.
In its third version (2020), AMIE is best in class in terms of rule mining speed and quality.

[1] https://github.com/lajus/amie

PhD objective 1: Handle tabular data

Current rule mining algorithms are limited to knowledge bases, or tabular data with less than 2 columns.

Idea

Extend the current exploration algorithm of AMIE to explore all join conditions in parallel.

Expected benefits

Generalization to all kinds of tabular datasets: Nhtsa, Legifrance, has-sante, texastribune... Being able to mine rules such that:
IF an employee works in texas government as a data scientist THEN employee's annual salary increase matchs inflation rate.

PhD objective 2: Mine rules with numerical attributes

We want to be able to mine numerical comparisons on data such as $<, >, =$. This is challenging because the search space is infinite.

Idea (see vision paper[5])

- Starting out with comparisons between attributes of entities.
- Binary searches for finding thresholds for numerical attributes.
- Bucketing.

Expected benefits

Being able to mine rules such that:
If candidate age $>$ 50 Then final acceptance ratio is $<$ 10%.

PhD objective 3 : Mine rules with negations

Knowledge bases or tabular data do not contain negative information. In addition, due to the open world assumption we cannot infer that absent statements are negative statements.

Idea

Adapt more methods [9] [7] that estimate when an absent statement is negative. When can we detect that the absence of information means something specific ?

Expected benefits

Being able to mine rules such that:
IF employee ethnicity is majority THEN there are NO decrease NOR increase in acceptance ratio.

PhD proposal François Amat

Goals

Mine rules such as : **If candidate age $>$ 50 and job is data scientist in Europe Then there are no decreases in acceptation ratio**.

- On tabular data
- With numerical attributes
- With negations

Numerous applications

AI & Data for Business

Help expert users to understand:

- **Reasons**: Deaths in car accident are linked with the part **seat belt**.
- **Risk management**: Using the **DOL** argument have High risk for fiscal issues.

AI & Data for Society

Being able to check:

- **Compliance**: If candidate age $>$ 50 Then final acceptance ratio is NOT lower than other candidates.
- **Dependencies**: Company name does NOT implies high reimbursement rate.

References I

A. Adadi and M. Berrada.
Peeking inside the black-box: A survey on explainable artificial intelligence (xai).
IEEE Access, 6:52138–52160, 2018.

Valérie Beaudouin, Isabelle Bloch, David Bounie, Stéphan Clémençon, Florence d'Alché Buc, James Eagan, Winston Maxwell, Pavlo Mozharovskyi, and Jayneel Parekh.
Flexible and context-specific ai explainability: A multidisciplinary approach, 2020.

Finale Doshi-Velez and Been Kim.
Towards a rigorous science of interpretable machine learning, 2017.

References II

F. K. Došilović, M. Brčić, and N. Hlupić.
Explainable artificial intelligence: A survey.
In *2018 41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, pages 0210–0215, 2018.

Luis Galárraga and Fabian M. Suchanek.
Towards a Numerical Rule Mining Language.
In *AKBC workshop*, Montreal, Canada, December 2014.

Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi.
A survey of methods for explaining black box models.
ACM Comput. Surv., 51(5), August 2018.

References III

Jonathan Lajus, Luis Galárraga, and Fabian Suchanek.
Fast and Exact Rule Mining with AMIE 3.
In *ESWC 2020: The Semantic Web*, pages 36–52, Virtual Event, Greece, June 2020.

Breiman Leo.
Classification and regression trees / Leo Breiman,... Jerome H. Friedman,... Richard A. Olshen... [et al.].
«The »Wadsworth and Brook-Cole statistics-probability series. Chapman & Hall/CRC, Boca Raton [etc, C 1993.

Stefano Ortona, Venkata Vamsikrishna Meduri, and Paolo Papotti.
Rudik: Rule discovery in knowledge bases.
Proc. VLDB Endow., 11(12):1946–1949, 2018.

References IV

Fabian Suchanek.
The Need to Move beyond Triples, 2020.

Fabian M. Suchanek, M. Sozio, and G. Weikum.
Sofie: a self-organizing framework for information extraction.
In *WWW '09*, 2009.

Berk Ustun and Cynthia Rudin.
Supersparse linear integer models for optimized medical scoring systems.
Machine Learning, 102, 03 2016.

References V

Hongyu Yang, Cynthia Rudin, and Margo Seltzer.
Scalable Bayesian rule lists.
In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 3921–3930. PMLR, 06–11 Aug 2017.

Backup slides

Database Approaches

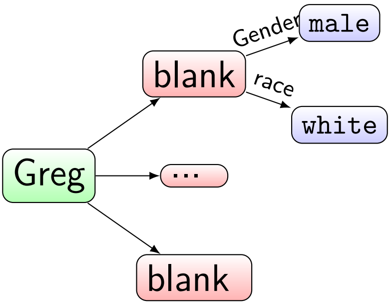
- **Key contraints**
  Example: "Employee id" 1:1 with "Employee name"
- **Foreing key constraints**
  Example: the tables *Employee_office*,*Employe* are linked.
- **Association rules**
  Example: If Salary is over $40k Then employement status is "Full time"

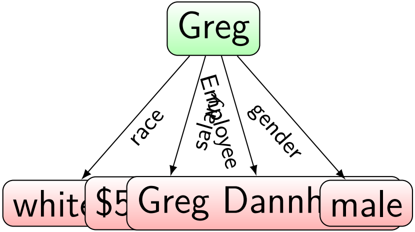
Explainable and causation

There is considerable debate about the meanings of the terms "explainable" and "interpretable", and what constitutes "causation" [1] [3] [2] [4] [6].
In our work, we aim at *interpretability* in the following sense: We want to provide a meaning for the results of a model in terms that are understandable to humans [3].

Tabular to kb (wip)

Employee	Gender	Race	Hours per week	Salary
Greg Dannheim	Male	White	40	$50,078





AMIE performances

Table 6: Performances and output of Ontological Pathfinding (OP), RuDiK and AMIE 3. *: rules with support ≥ 100 and CWA confidence ≥ 0.1.

Dataset	System	Rules	Runtime
Yago2s	OP (their candidates)	429 (52*)	18min 50s
	OP (our candidates)	1 348 (96*)	3h 20min
	RuDiK	17	37min 30s
	AMIE 3	97	**1min 50s**
	AMIE 3 (support=1)	1 596	7min 6s
DBpedia 3.8	OP (our candidates)	7 714 (220*)	> 45h
	RuDiK	650	12h 10min
	RuDiK + types	650	11h 52min
	AMIE 3	5 084	**7min 52s**
	AMIE 3 (support=1)	132 958	32min 57s
Wikidata 2019	OP (our candidates)	15 999 (326*)	> 48h
	RuDiK	1 145	23h
	AMIE 3	8 662	**16h 43min**

Definitions

Open world assumption

We only know what we have in the database.
Example: If employee.maritalStatus = Null then it does not mean that the employee is not maried.

Knowledge base

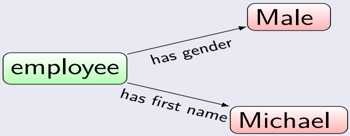


Figure: knowledge base example

They can only store binary predicate, called relation.

PhD objective 4: Improve the predictive power

Predict a new statement is not a trivial task, even if we can mine all rules efficiently. Indeed, rules have to be combined to arrive at new statements and gauge their probability.

Idea

Use of logical reasoning [11] and probabilistic methods such as Markov Logic Networks

Expected benefits

Being able to predicting a new statement such as : Michael has a professional cell phone because he has "director" in his title.

PhD objective 5: Mine rules with Meta-relations

Current ILP does not take into account statements about statements.

Idea

Collaboration with the NoRDF project [10].

Expected benefits

Being able to use statements about statements in rule mining would allow to have better rules. Indeed, this would allow distinguishing statements that are beliefs, refused, old... For instance, "In this company, all executives had the same gender until 2017"

Black box models are great but not suited for all industries.

We want to allow domain experts to use AI for critical tasks.

Black box models make predictions based on input data.
Examples: Deep Learning models, Random Forests.

$$\text{data} \longrightarrow \boxed{\text{Black box}} \longrightarrow \text{prediction}$$

Black box models are great for making accurate predictions, but their output **cannot be explained**. Critical tasks in **security**, **health** or **justice** cannot be operated by black box predictions. Indeed, due to liabilities, requirements, understanding why a prediction and therefore why an action is made, **must be justified**.