Mapping the challenges and opportunities of CBR for eXplainable AI

Belén Díaz Agudo

27th International Conference on Case-Based Reasoning
September 8 – 12, 2019, in Otzenhausen, Germany
https://iccbr2019.com/

My research

- Knowledge Intensive CBR
- Applications
  - Recommender systems
  - Videogames, storyplot and behaviour generation
  - Learning systems
- Explanations and CBR
  - From KI CBR to KI explanations
Outline

1. What is XAI and XCBR?
   - Motivations and definitions
   - What, When and How to explain?

2. Is XCBR a key to open the black box of AI systems?

3. Challenges and future

References

Christoph Molnar. April 2019

AAAi 2019 XAI tutorial
https://xaitutorial2019.github.io/

Explanation and justification in machine learning: A survey. IJCAI-17 XAI Workshop
Biran, O., et al. (2017)


References at these slides (to share)
Is XCBR a key to open the black box of AI systems?

Challenges and Future

Part 1

Explainability was ranked by physicians as the most desirable feature of a clinical decision support system (Teach RL, Shortliffe EH. 1981)

CHEF (Hammond, 1989) explained the failures in recipes

Since MYCIN (1994) most KBSs provide some form of explanation

XAI is not new
Research on Explanations and CBR

Belen Díaz Agudo – ICCBR 2019

Timeline

P. Cunningham, D. Doyle, J. Loughrey, An Evaluation of the Usefulness of Case-Based Explanation, ICCBR 2003
D. Doyle, P. Cunningham, D. G. Bridge, Yusof Rahman: Explanation Oriented Retrieval. ECCBR 2004
Thomas Roth-Berghofer: Explanations and Case-Based Reasoning: Foundational Issues. ECCBR 2004

1988
Early papers

1994 1996

2004

2018 2019

Algorithmic decision-making on the rise

Critical systems

Can AI systems be more **objective** than humans?

Bias, ethical and transparency of AI in the media

**Trust**

Sam Corbett-Davies, Emma Pierson, Avi Feller and Sharad Goel (2016). "A computer program used for bail and sentencing decisions was labeled biased against blacks. It's actually not that clear". *The Washington Post.*
Motivating Examples

**Criminal Justice**
- People wrongly denied parole
- Recidivism prediction
- Unfair Police dispatch

Glenn Rodríguez (inmate in New York) was denied parole last year despite having a nearly perfect record of rehabilitation. The reason? A high score in recidivism prediction from a computer system called **Compas**.

**Motivating Examples**

**Finances**
- Credit scoring, loan approval, mortgages
- Insurance quotes
- Biases

"Artificial intelligence could revolutionise the insurance industry but it may also allow clients to calculate if they need protection."

Financial Times looks at the role that AI plays in the world of insurance, and how Aerobotics is using machine learning algorithms to help insurers improve efficiency.

Read the full press release →

Source: https://www.ft.com/content/e07cee0c-3949-11e7-821...
Ethical questions

Do we trust the system?

Motivating Examples


[Caruana et al. 2015] "Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission."

Explanation

Definition of explanation

1. the act or process of explaining
2. something that explains // gave no explanation

Why did not the treatment work on the patient?
Why was my loan rejected?
Why is it raining in Spain?

Why do they put holes in crackers?
What is this for?
How does it work?
How to program in Java?
Humans use imperfect, incomplete and common sense knowledge. Explanations are interactions between the explainer and the explainee.


Learning
Teach the user about the domain

Conceptualization
Meaning of concepts

Transparency
Explain how the system reached the answer

Justification (post-hoc)
Explain why the answer is a good answer

Relevance

What to explain?
Explanation Goals

- Types of explanations for Expert Systems [Spicker 1991]
  - Conceptual explanations
  - Why explanations
  - How explanations
  - Purpose explanations
  - Cognitive explanations

Belen Díaz Agudo – ICCBR 2019
Understanding the reasons or causes behind AI reasoning

Increase Trust
- Get insights into the model
- Add transparency
- Reduce vulnerability

XAI goals

When do we explain?
The needs for explanations
- Increase trust
- Curiosity
- Monitoring
- Learning


Belen Díaz Agudo – ICCBR 2019
AI Explainability / Interpretability

The degree to which a human can understand the cause of a decision
The degree to which a human can consistently predict the model’s result

Explanation in Artificial Intelligence: Insights from the Social Sciences. Tim Miller 2017


“XAI aims to create a suite of techniques that produce more explainable models, while maintaining a high level of performance, accuracy and precision”
Interpretability

Interpretable models
Linear regression
Logistic regression
Decision tree
GLM (Generalized Linear Model)
GAM (Generalized additive models)
Naive Bayes
K Nearest Neighbours
Rules
Partial dependence plots
Attribution or influence scores
Sensitivity Analysis
Prototype Selection
Activation Maximization


Black box ML models
Neural Networks
DNN Deep Neural Networks
SVM Support Vector Machines
Random Forest
Genetic algorithms
Tree ensemble (TE)

Transparent Design vs Post-hoc Explanation

**Transparent design** reveals *how* a model functions.

Model-specific explanation methods

Explanations

Model-agnostic explanation methods

Post-hoc Explanation explains *why* a black-box model behaved that way.

Justifications

[AAAI 2019 tutorial on explainable AI](https://aaai2019.github.io/)

How to explain?
ALIBI

ALIBI is an open source Python library aimed at machine learning model inspection and interpretation. The initial focus on the library is on black-box, instance based model explanations.

Goals

- Provide high-quality reference implementations of black-box ML model explanation algorithms
- Define a consistent API for interpretable ML methods
- Support multiple use cases (e.g. tabular, text and image data classification, regression)
- Implement the latest model explanation, concept drift, algorithmic bias detection and other ML model monitoring and interpretation methods

HOW to explain?

Explanation Algorithms

LIME (Local Interpretable Model-Agnostic Explanations)  "Why Should I Trust You?": Explaining the Predictions of Any Classifier  (Ribeiro, Singh and Guestrin 2016)

ANCHOR (Ribeiro, Singh, and Guestrin 2018)

Surrogate Model
Extracting tree-structured representations of trained networks  (Craven and Shavlik 1995)

SHAP (SHapley Additive exPlanations)  A Unified Approach to Interpreting Model Predictions,  (Lundberg and Lee 2017)


Contrastive Explanations with Local Foil Trees  (Jasper van der Waa et al 2018)

STREAK  Streaming Weak Submodularity: Interpreting Neural Networks on the Fly  (Elenberg et al 2017)


Layer-wise Relevance Propagation (LRP)  Layer-wise Relevance Propagation for Neural Networks with Local Renormalization Layers.  (Binder et al. 2016)

AUTO-ENCODER  Deep Learning for Case-Based Reasoning through Prototypes: A Neural Network that Explains Its Predictions.  (Li, Liu, Chen and Cynthia Rudin 2017)

RxREN  Trust scores

DeepLIFT

QII
HOW to explain?
Explanation Algorithms

**LIME** (Local Interpretable Model-Agnostic Explanations)
"Why Should I Trust You?": Explaining the Predictions of Any Classifier (Ribeiro, Singh and Guestrin 2016)

**ANCHOR** (Ribeiro, Singh, and Guestrin 2018)

**RETAIIN**

**LRP**
Layer-wise Relevance Propagation for Neural Networks with Local Renormalization Layers. (Binder et al. 2016)

**STREAK**
Streaming Weak Submodularity: Interpreting Neural Networks on the Fly (Elenberg et al 2017)

**SHAP**
A Unified Approach to Interpreting Model Predictions, (Lundberg and Lee 2017)

**DeepLIFT**
Surrogate Model
Extracting tree-structured representations of trained networks (Craven and Shavilk 1995)

**QII**
Contrastive Explanations with Local Foil Trees (Jasper van der Waa et al 2018)

**RxREN**
Trust scores

**DeepLIFT**
Surrogate Model

**Training local surrogate models Perturbations**

**Perturbing** the input query and see **how the predictions change**
Prospector – PDP, AGN, TAB

- Introduce *random perturbations* on input values to understand to which extent every feature impact the prediction using PDPs.
- The input is changed *one variable at a time*.

Contrastive explanation model (CEM)

**Contrastive Explanations with Local Foil Trees**

1. Data point → Model → Class A
   - Class A as "Fact"

2. Class distribution data point
   - A, B, C

3. Sample local dataset around data point

4. Train Decision Tree on foil class

5. Data point
   - Locate the fact-leaf

6. Data point
   - Locate the foil-leaf (closest leaf with foil)

7. Data point
   - Retrieve different decision nodes

8. "This data point is classified as A instead of B, because <feature x> is less than <threshold T>."

Select or create particular instances that are used to explain

Representative prototypes vs criticisms (exceptions)

k-nearest neighbors

Influential instances

Counterfactuals

Example based explanation methods

Algorithm that finds prototypes and criticisms

Find weaknesses of a ML model

Prototypes

Criticisms

A prototype is a data instance that is representative of all the data.

A criticism is a data instance that is not well represented by the set of prototypes.

Protoypes and criticisms

Kim, Been, Rajiv Khanna, and Oluwasanmi O. Koyejo. “Examples are not enough, learn to criticize! Criticism for interpretability.” (2016)
Peter applies for a loan rejected by the banking software. **Why?**

Counterfactuals: “hypothetical reality that contradicts the observed facts”

What is the smallest change to the features (income, number of credit cards, age, ...) that would change the prediction from rejected to approved?

**Counterfactual explanations**

If Peter had fewer credit cards and had not defaulted on a loan 5 years ago, he would get the loan.

If Peter would earn 10,000 Euro more per year, he would get the loan.

Not transparency

Peter doesn’t know the real reasons for the rejection

**Counterfactual explanations**


Which features are sufficient to anchor a prediction, i.e. changing the other features cannot change the prediction?
XAI and XCBR

Is XCBR a key to open the black box of AI systems?

Challenges and Future

Part 2

Does CBR have a role?

Model agnostic (post hoc) explanations using CBR

CBR surrogates

Example based explanations

Anchors/counterfactuals

Examples/prototypes/criticisms

Use CBR for the interpretable model

Do we have the training data?
CBR surrogates (challenge)

1. D = Perturb query
2. D+ = Predicted class labels for each member of D
3. Retrieve members of D+ weighted with similarity to Query

Similar cases with the same or different classes

Belen Díaz Agudo – ICCBR 2019
**CBR surrogates (research questions)**

What is the Perturbation method?
How many cases to produce for D?
Which cases?
- Nearest-like/ unlike-neighbours
- Create a case base around the decision boundary

[Doyle et al ECCBR 2004]
[I. Watson XCBR Workshop. 2018]

---

**Explanation oriented retrieval**

[Doyle et al 2004] Explanation Oriented Retrieval ECCBR 2004

*Fig. 1. A nearest neighbour example where case EC would be a better explanation for the decision on query case Q than the nearest neighbour NN; case NUN is the nearest unlike neighbour*
What kinds of case-based explanations can we produce from the case base?

- Most similar neighbor(s) and contrasting them with unlike neighbors
- Counterfactuals
- A fortiori cases

Can we improve this incorporating external knowledge in the different steps?

What if we have the training data for the black box model?
CBR surrogates (research questions)

What if we have the **training data** for the black box model?

Use the training data as the case base

**Apply learning strategies**
- Learning a similarity metric on the cases
- Complex perturbations methods

---

**Combination of two techniques NN + CBR**
- Run in separate modules with a **common dataset**
- Separate labor: predictions + interpretability
- Sequence: Features weights generated by the NN are used for retrieval in the CBR system

---

Mark T Keane and Eoin M Kenny. How Case-Based Reasoning Explains Neural Networks: a theoretical analysis of XAI using post-hoc explanation by example from a survey of ANN-CBR twin systems. ICCBR 2019


---

**ANN-CBR Twins**
Other CBR Twins

CBR confidence as a basis for confidence in Black Box Systems. L. Gates, C. Kisby and D. Leake. *ICCBR 2019*

**COBB (Case-Based cOnfidence for Black Box)**

A Case-Based Explanation System for Black-Box Systems. C. Nugent, P. Cunningham. *Artificial Intelligence Review 2005*

Investigating Textual Case-Based XAI. R. Weber, A. Johs, J. Li and K. Huang. *ICCBR 2018*

Does CBR have a role?

What if we have additional knowledge?

**SWALE** style XPs (Kass and Leake, 1988)
PARIS Plan Abstraction and Refinement in an Integrated System (Bergmann et al., 1993)
DIRAS (Armengol et al., 2000)

Retrieve and Reuse Human explanations

Dataset: <D, S, E>
E:[Textual] explanations from an expert

Case base of [human] explanations

Learn, retrieve and reuse explanation cases
SWALE uses CBR for the **creative** generation of explanations of **anomalies**. Its explanation process is based on the retrieval and application of **cases storing prior explanations**, called **explanation patterns** from its memory.

Similarity assessment among situations is a key issue.

SWALE needs a high amount of **domain knowledge** in the form of relations between concepts and explanation patterns.


https://www.cs.indiana.edu/~leake/projects/swale/

Why did Swale die?

**Remindings from Swale’s Death**

- A vet: “This sounds like an aneurysm. I’ve seen this sort of thing before.”
- A Yale AI lab student: “This sounds like the death of Jim Fixx” (a runner, who died of a heart-attack when in peak condition.)
- Another lab student: “Swale was a young superstar like Janis Joplin. Maybe he died of a drug overdose.” Could this apply to Swale?

*Case-Based Explanation!*
*Case-Based Creativity!*
Recent XAI techniques reusing human generated explanations

Argumentation, explanation as cooperative conversation, and dialog

Context and Personalization

Inside and outside the CBR community


Data explanation with CBR. Belen Diaz-Agudo, Juan Recio-Garcia & Guillermo Jimenez-Diaz. In XCBR Workshop ICCBR 2018.


**Case base of explanations**

Belen Díaz Agudo – ICCBR 2019

**XCBR: Explanations and CBR**

**Model agnostic**

![Diagram showing a model agnostic approach to CBR](image)

**Model Specific**

![Diagram showing a model specific approach to CBR](image)

Is CBR always an interpretable and transparent model?

Belen Díaz Agudo – ICCBR 2019
Explaining CBR

When does a CBR system need explanations?

“A solution to a new problem can be explained simply by presenting the case that was adapted to create the solution” [Riesbeck 1988]

“CBR can present cases to the users to provide compelling support for the system conclusions” [Leake 1996]

- ... but cases are not always good explanations
  - Can you represent an instance in a humanly understandable way?
  - Can you can ‘interpret’ a single instance in the dataset?

Belen Díaz Agudo – ICCBR 2019

Visual Explanations

Explainable artificial intelligence for breast cancer:
A visual case-based reasoning approach.
Lamy J Guezennecc, G Bouaud J et al.
Artificial Intelligence in Medicine (2019)
Explaining CBR

“... but cases are not always good explanations
  - Can you represent an instance in a humanly understandable way?
  - Can you ‘interpret’ a single instance in the dataset?
  - Cases do not provide information about how they were selected and adapted
  - CBR struggles to explain the results concerning the domain knowledge

CBR can present cases to the users to provide compelling support for the system conclusions” [Leake 1996]

Visual explanations

Transparency

Post hoc similarity explanations

Explaining similarity in CBR. ECCBR 2004
E. Armengol, S. Ontañon and E. Plaza.

A. Sánchez-Ruiz, S. Ontañón:
Explanations as part of the reasoning process

Investigating the solution space for online iterative explanation in goal reasoning agents. AI Commun 31 (2018). Christine Task, Mark A. Wilson, Matthew Molineaux, David W. Aha

Using explanations to provide transparency during trust-guided behavior adaptation. AI Commun. 30 (2017). Michael W. Floyd, David W. Aha


Explanation-based similarity for case retrieval and adaptation. MoCAS (Pews and Wess, 1993)

Agnar Aadmodt. Explanation driven CBR. 1994

Anders Kofod-Petersen, Jörg Cassens, Agnar Aamodt: Explanatory Capabilities in the CREEK Knowledge-Intensive Case-Based Reasoner. 2008

What knowledge do we need to explain CBR?

- Rich knowledge models to produce context dependent semantic explanations

Knowledge elicitation and formalisation for context and explanation-aware computing with case-based recommender systems. Christian S. Sauer (Phd) 2016

Thomas Roth-Berghofer, Jörg Cassens: Mapping Goals and Kinds of Explanations to the Knowledge Containers of Case-Based Reasoning Systems. ICCBR 2005
Explanable AI: The new 42?

**INPUTS**

**OUTPUTS**

**CONTEXT**

**EXPLANATION**

How a certain output depends on the inputs?


---

**Collaborative filtering**

Matrix factorization

**Content based**

Interpretable

Explicit user profile

---

Explanations in recommender systems

Belen Díaz Agudo – ICCBR 2019
Explanations in recommender systems

Jon Herlocker et al 2001. Explaining Collaborative Filtering Recommendations

Derek Bridge and Kevin Dunleavy: If you liked Herlocker et al.’s explanations paper, then you might like this paper too, in *Workshop of the ACM Conference on Recommender Systems*, 2014.


| Personalized case-based explanation of matrix factorization recommendations |
| Jorro et al. ICCBR 2019 |

| Explanation of recommenders using Formal Concept Analysis |
| Belen Diaz-Agudo, et al. ICCBR 2019 |


Open research questions
What Is a Good Explanation?

- Evaluative
- Inclusive
- Truthful
- Selective
- Contrastive
- Short
- Stable
- Certain
- Fidel
- Responsive
- Useful
- Understandable
- Probable
- Comprehensible
- Consistent with prior beliefs
- Personalized
- Focus on the abnormal/exceptions
- Novel
- Representativeness

Difficult to evaluate

Evaluation

Belen Díaz Agudo – ICCBR 2019

Part 3

What is XAI?

Is XCBR a key to open the black box of AI systems?

Challenges and Future

Belen Díaz Agudo – ICCBR 2019
Open Research Questions

- There is no agreement on what an explanation is
- There is not a formalism for explanations
- There is no work that seriously addresses the problem of quantifying the grade of comprehensibility of an explanation for humans
- Is it possible to join local explanations to build a globally interpretable model?
- What happens when black box make decision in presence of latent features?
- What if there is a cost for querying a black box?

Key Questions for XCBR (and XAI)

- Key questions:
  - What does the user want/need to know?
  - What is the explanatory context?
  - What explanations have been seen before?
  - What has the system learned about the user?
  - What has the user learned about the system?
**CBR has a key role in the general goal of having an AI more transparent, responsible and accountable**

- Memory of cases of explanation
- Reuse *human explanations*
- Case-Based Surrogates and twins
- Explanation by example

**CBR to open the black box AI systems**

---

**Is CBR a key?**

Belen Díaz Agudo – ICCBR 2019

---

**Challenges for the CBR community**

**Example based explanations**
- Sensible data/ privacy/confidenciality
  - Pertubations on the real data? Data visualization

**CBR surrogates or twins**: justifications not explanations
- Faithfulness to the underpinning model

**Human explanatory experience**
- Personalization and context
- Creativity, provenance, learning, narrative, conversations

**Measuring explanation effectiveness**
- Human oriented explanations? Human studies to evaluate explanations

**Visualize XCBR in the XAI community**
- Be aware of work in CBR reported at other events

Belen Díaz Agudo – ICCBR 2019
**Take home message**

1. Explanation is a reasoning process itself
   - Can it be a CBR process? Similar situations have similar explanations?

2. XCBR is not only example based explanations
   - Use other knowledge Experience

3. Experience of the human being explainer
   - Provenance Learning

4. XCBR as a systematic approach for XAI

5. Important aspects from the early papers
   - Creativity and narrative Conversations

6. Explanations for People?
   - Human friendly explanations Interfaces and personalization Evaluation with humans

**The role of the experience in the explanation process**
Belen Díaz Agudo – ICCBR 2019

---

**The Big Question: How Can We Make Systems Explain Themselves Better?**

Any question?