Using Formal Concept Analysis to Explain Black Box Deep Learning Classification Models

Amit Sangroya, C. Anantaram, Mrinal Rawat, and Mouli Rastogi

TCS Innovation Labs, India
{amit.sangroya, c.anantaram, rawat.mrinal, mouli.r}@tcs.com

Abstract. Recently many machine learning based AI systems have been designed as black boxes. These are the systems that hide the internal logic from the users. Lack of transparency in decision making limits their use in various real world applications. In this paper, we propose a framework that utilizes formal concept analysis to explain AI models. We use classification analysis to study abnormalities in the data which is further used to explain the outcome of machine learning model. The ML method used to demonstrate the ideas is two class classification problem. We validate the proposed framework using a real world machine learning task: diabetes prediction. Our results show that using a formal concept analysis approach can result in better explanations.

1 Introduction

Deep learning techniques have improved the state of the art results in various areas such as natural language processing, computer vision, image processing etc. The area is growing at such a fast pace that everyday a new model is being discovered that improves the state of art rapidly. One of the area that is still under studied is related to the use of these models in real-world such that the outcome can be explained effectively. For instance, if a critical AI (Artificial Intelligence) system such as medical diagnosis only tells whether a patient has a certain disease or not without providing explicit reasons, the users can hardly be convinced of the judgment. Therefore, the ability to explain the decision is an important aspect of any AI system particular natural language processing (NLP) based system.

Recently, lots of works have been done to solve natural language processing research problems such as text classification, sentiment analysis, question answering etc. However, there are very few attempts to explore explainability of such applications. Relational data is usually described by objects and their attributes. Particularly, structure of data is defined by dependencies between the attributes. Explanation consists of performing an exception and transformation analysis to validate the outcome of a ML model. In this paper, our approach to explanation generation is via using formal concept analysis, a conceptually different perspective from existing approaches. A central goal of this research is to build a general purpose or domain-independent framework for interpreting classification outcome of deep learning models, rather than just a single problem in a particular domain. In summary, our contributions in this work are as follows:

– We propose a formal concept analysis based approach in order to generate explanations for the outcomes.
– Furthermore, we show the effectiveness of our method on a real world data set i.e. diabetes prediction.

2 Framework

In this paper, we approach the explanation generation problem from a different perspective – one based on formal concept analysis (FCA). We propose a general concept lattice theory based framework for explanation generation, where given an outcome O of a deep learning model and a domain ontology, the goal is to identify an explanation that can point the user to the prominent feature set f for a certain outcome. We use diabetes classification as an example to evaluate the framework where we model two situations: One where outcome of deep learning black box model and outcome of FCA based classification directly matches and one where it does not match. Further, we present an algorithm, implemented for FCA, that computes such similarities and evaluate its performance experimentally. In addition to providing an alternative approach to solve the explanation generation problem, our approach has the merit of being more generalizable to other problems beyond classification problems as long as they can be modeled using a FCA based concept lattice.

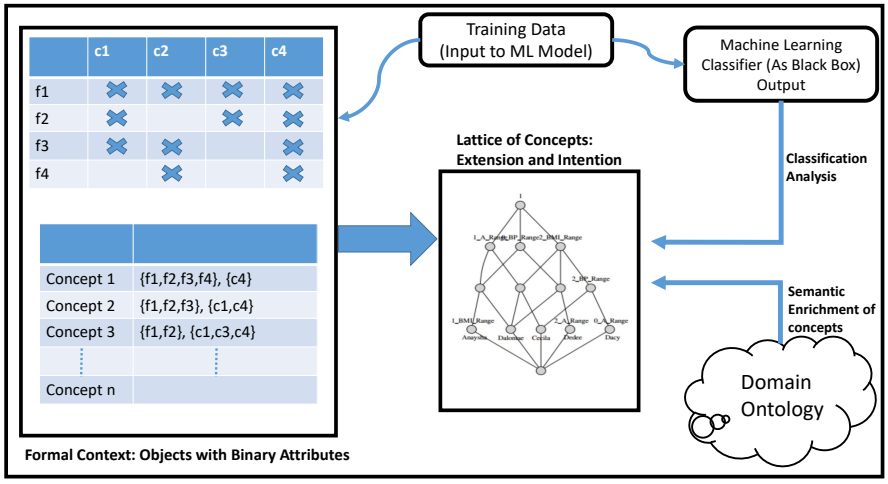

Fig. 1: Overview of the Proposed Framework

Algorithm 1 Explanation of Black Box ML model

Input: M, $c0$, $c1$, $samples$ ▷ M: ML Model; $c0$: lattice of class zero; $c1$: lattice of class one
Output: E ▷ Explanations
procedure PREDICT_FCA($c0$, $c1$, s_i)
    $P \leftarrow \emptyset$ ▷ Prediction
    $class_0\_lattice \leftarrow c0.lattice$
    $class_1\_lattice \leftarrow c1.lattice$
    $s_i_lattice \leftarrow$LOAD_FCA(s)
    for extent, intent $e_j, i_j \in s_i\_lattice$ **do**
        for extent, intent $e_k, i_k \in class_0\_lattice$ **do**
            if $i_k.issubset(i_j)$ **then**
                $P \leftarrow 1$)
            end if
        end for
        for extent, intent $e_k, i_k \in class_1\_lattice$ **do**
            if $i_k.issubset(i_j)$ **then**
                $P \leftarrow 0$
            end if
        end for
    end for
    return P
end procedure
procedure EXPLANATIONGENERATOR(S, D)
    $P_{ML} \leftarrow \emptyset$ ▷ ML Predictions
    $P_{FCA} \leftarrow \emptyset$ ▷ FCA Predictions
    $E \leftarrow \emptyset$
    for sample $s_i \in samples$ **do**
        $p \leftarrow M$.predict(s_i)
        if $p > 0.5$ **then**
            P_{ML}.add(1);
        else
            P_{ML}.add(0);
        end if
        P_{FCA}.add(PREDICT_FCA(s_i))
        for feature $f_j \in s_i$ **do**
            $f_j \leftarrow$ MODIFY(f_j)
            $P \leftarrow$ PREDICT_FCA(s_i)
            if $P_{ML_i} == P_{FCA_i}$ **then**
                if $P \neq P_{FCA_i}$ **then** E.add(Feature j Sample i may be responsible to classify as 0);
                else E.add(Feature j Sample i may be responsible to classify as 1);
                end if
            else
                if $P \neq P_{FCA_i}$ **then** E.add(Feature j Sample i may be responsible to classify as 1);
                else E.add(Feature j Sample i may be responsible to classify as 0);
                end if
            end if
        end for
    end for
    return E
end procedure

2.1 Formal Concept Analysis

The fundamental fact underlying FCA is the representability of complete lattices by ordered sets of their meet and join irreducibles. Since ordered sets of irreducibles are naturally represented by binary matrices, this makes it possible to apply certain aspects of the lattice theory to the analysis of data given by object-attribute matrices.

Formal Concept Analysis starts with a formal context (G, M, I) where G denotes an ordered set of objects, M a set of attributes, or items, and I $\subseteq$ G $\times$ M a binary relation between G and M [1]. The statement (g, m) $\in$ I, or gIm, means: "the object g has attribute m". Two operators $(\cdot)'$ define a Galois connection between the power sets (P(G),$\subseteq$) and (P(M),$\subseteq$), with A$\subseteq$G and B$\subseteq$M: $A' = \{m \in M | \forall g \in A : gIm\}$ and $B' = \{g \in G | \forall m \in B : gIm\}$. A pair (A, B), such that A$'$ = B and B$'$ = A, is called a formal concept, where A is called the extent and B the intent of the concept (A, B). The set of all formal concepts of (G, M, I) created by a partial order relation $\leq$, is a subconcept-superconcept hierarchy and is called the concept lattice $\mathcal{L}$.

2.2 Implication Rules

Implication rules S $\Longrightarrow$ T, where S,T $\subseteq$ M holds in context (G,M,I) if S$' \subseteq$ T$'$ i.e., each object having all attributes from S also has all attributes from T. These rules are significant as they expresses the underlying knowledge of interaction among attributes and moreover, also contains statistical values like support and confidence.

2.3 Classification Analysis

Classification analysis is done to predict the category of new as well as existing objects. This is carried out by defining a target attribute in the dataset, generating concept lattices for each value of the target attribute and then comparing new/existing object's attributes with the intents of the concept lattice for each category. In this analysis, a query asking for object details is posed. Lattice structures corresponding to each target value is stored in the memory. Moreover, if an intent *i* of a lattice contains some intent *j* of another lattice, then intent *j* is not considered in the analysis. At the run time, attributes set matching of the new/existing object is done with each of the lattices in the memory. If there is only one lattice *L* whose some concept's intent contains the intent of new/existing object, then the corresponding category is assigned to that object otherwise the result "*not determine*" is declared.

2.4 Semantic Enrichment using Domain Ontology

Ontology is the formal specification of concepts and relationships for a particular domain (e.g. in the domain of finance, US-GAAP is widely used ontology). Ontology has a formal semantic representation that can be used for sharing and reusing knowledge and data. We have downloaded ontology for diabetes from bioportal.bioontology.org/ontologies/DIAB. In the next step, these concepts and relationships are subsequently coded in the Web Ontology Language (OWL) with Protege.

Table 1: Example of Diabetes Ontology

Subject	**Predicate**	**Object**
type 2 diabetes mellitus	has_exact_synonym	type II diabetes mellitus
type 2 diabetes mellitus	has_exact_synonym	non-insulin-dependent diabetes mellitus
type 2 diabetes mellitus	has_exact_synonym	NIDDM
type 2 diabetes mellitus	is_subClassOf	diabetes mellitus
diastolic blood pressure	has_low_range	< 70
diastolic blood pressure	has_high_range	> 100
body mass index	has_normal_range	< 23

This ontology (stored as a Resource Description Framework graph) stores the concepts of the domain and their relationships with a `<subject-predicate-object>` structure for each of the concepts. For instance, Table 1 shows an example of diabetes ontology. Here, concepts like diabetes etc. are defined along with concept relationships and synonyms. Additionally, ontology also define the categorical partitioning of diabetic attributes based on medical experts opinion. For example, ontology suggests the normal, low and high ranges for blood pressure. This ontology also assists in deriving implication rules which assists in classification analysis through FCA.

3 Results

The data for diabetes prediction is taken from www.kaggle.com/uciml/pima-indians-diabetes-database. The datasets consist of several medical predictor (independent) variables and one target (dependent) variable, Outcome. Independent variables include the number of pregnancies the patient has had, their BMI, insulin level, age, and so on.

Data pre-processing involves removing missing/invalid values. Thereafter, we enrich the data using a domain ontology. This involves defining ranges for the records and also building concept hierarchy. Thereafter, we build a ML model to classify if a certain object has diabetes or not. At the same time, we also use FCA approach to classify the same set of objects. Note that the objective of using FCA based classification was just to explain the outcome of ML model, which has been used as a black box. Results are summarized in Table 2.

Table 2: Results using FCA and ML Model

	ML Model	**FCA**
Accuracy	70%	73%
Precision	77%	72%
Recall	63%	90%

3.1 Classification using ML Approach

We used a LSTM based deep neural network based binary classification to train on the processed dataset. Number of training samples were 540 and testing samples were 150. We used all 11 features available in the data such as BMI, Blood Pressure, Insulin etc. The test accuracy of diabetes classification was 70% (See Table 2). Interestingly, accuracy of FCA approach was better. This can be due to the fact that size of dataset was not very large. It might be possible that on a larger dataset ML model might perform better. However, the scope of this work was never to compare the accuracy of two approaches, but to use FCA based lattice theory to explain the output of black box ML model. The explanation of the outcomes was generated using FCA model as explained in the next subsection.

3.2 Classification using FCA Approach

We divided the training data (the same data that was used in ML model) into two classes: diabetes and no diabetes. Then, we created two separate concept lattices for both classes as shown in Figure 2 and 3. For each sample in test set, we created its lattice alongwith extent and intent of each concept in the lattice. Thereafter, we compared the intents of concept in sample lattice with concept in both lattices (class lattices i.e. lattices of diabetes and no diabetes). The comparison is based on subset matching between sample lattice and class lattices. Wherever there is a match between lattices, that class is assigned as predicted class for the test sample.

3.3 Explaining the ML Model Outcome using FCA

We compared the outcomes of ML model and FCA based classification. We take each sample in the test set and we try to map to the feature set. The goal of explanation is to identify the feature which may be prominent to classify a given sample into a particular class. In order to achieve this, we change the features and observe the outcome with modified features. If the outcome with modified features change (i.e. changing a feature fi, leads to change in Outcome Oj), we can assert that fi is responsible for the outcome (See Algorithm 1 for details).

Table 3 shows the identified feature set for two classes. It shows the relative importance of each feature for identifying a sample into diabetes or no diabetes. In the scope of current work, we present the results with individual features only. Similar experiments can be performed to compute the feature sets as well. As we observe, *Age* is least important feature for an outcome of diabetes class, whereas *Blood Pressure* is most important feature. Similarly, for an outcome to be in non-diabetes class *BMI* is the most prominent feature.

Outcome (Based on the features and Implication rules): Aarav doesn't have diabetes.

In order to qualitatively evaluate the results, we identified implication rules from the training data as shown in Table 5. For a given test sample, we also used implication rules to validate the output. For Example: *Predict that whether Aarav has diabetes or not from his blood pressure, body mass index and age (See Table 4*).

Table 3: Feature Interpretation for two classes (Diabetes and Non Diabetes)

	Diabetes	**Non Diabetes**
Number of times of pregnancy — (# Preg)	15.6%	36.7%
Plasma glucose concentration every 2 hours in an oral glucose tolerance test — (Plasma)	13.2%	37.5%
diastolic blood pressure (mm Hg) —- (Diast BP)	16.4%	42.18%
triceps skin fold thickness (mm) —- (skin)	12.5%	41.4%
2-Hour serum insulin (mu U/ml) —- (insulin)	11.7%	43.7%
body mass index (weight in kg/(height in (mm)2) —- BMI	10.9%	45.3%
diabetes pedigree function — Pedigree	9.3%	43.7%
Age in years —- Age	5.4%	40.6%

Table 4: Classification Example using FCA

Person details	**Input from user**
Name	Aarav
Age	25, Age-range(2)
Blood Pressure	66, BP-range(1)
Body Mass Index	23.2, BMI-Range(2)

4 Related Work

Most machine learning model rely on validation set accuracy as a way of primary measurement of trust. However, there are limitations of these approaches in using models in a real world paradigm. Recognizing the importance of interpretations in assessing trust, various frameworks have been proposed that focus on interpretable models, especially for the medical domain [2,3,4]. While such models may be appropriate for some domains, they may not apply equally well to others. In the domain of computer vision, systems that rely on object detection to produce candidate alignments [5] or attention [6] are able to produce explanations for their predictions. However these models are constrained to specific neural network architectures. Our focus is on building general, model-agnostic explanations that can be applied to any classifier.

Another common approach for generating explanation is to build another model over the outcome of original model [7,8]. One limitation of this approach is that these models approximate the original model globally, thus interpreting outcomes at a fine grain level becomes a significant challenge. In order to interpret model at fine grain local level, LIME is a promising approach approximates the original model locally [9]. This approach is model and domain agnostic. However, using formal concept analysis based interpretation approach, the outcome can be interpreted with a sound theoretical basis.

Table 5: Implication rules

Rule	**# instances**
BP-range(2), Age-range(2) $\Longrightarrow$ Outcome(0)	226
BMI-range(1), BP-range(1) $\Longrightarrow$ Outcome(0)	128
BMI-Range(2), BP-Range(2) $\Longrightarrow$ Outcome(1)	63
Age-Range(1), BMI-Range(2), BP-Range(1) $\Longrightarrow$ Outcome(1)	41
BP-Range(0), Age-Range(2), BMI-Range(0) $\Longrightarrow$ Outcome(0)	95
BP-Range(0), Age-Range(2), BMI-Range(2) $\Longrightarrow$ Outcome(1)	86
BP-Range(1), Age-Range(1), BMI-Range(2) $\Longrightarrow$ Outcome(1)	178

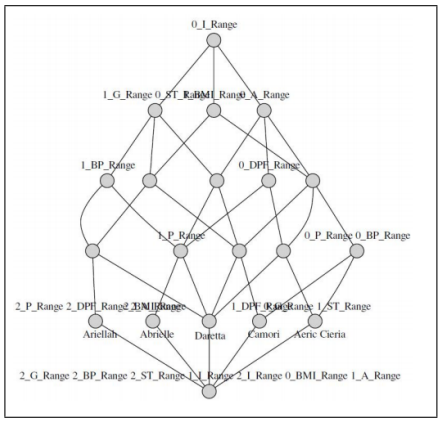

Fig. 2: No Diabetes

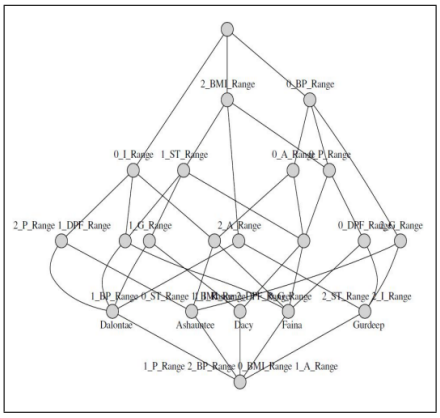

Fig. 3: Diabetes

5 Conclusion and Future Work

We considered Formal Concept Analysis in context of interpretation of machine learning models particularly focusing on classification and assuming that model to be explained is a black box model. The main attention was drawn to the lattice based classification analysis of attributes. We showed the significance using well known classification problem i.e. diabetes prediction. In this paper, we limited our experiments to two class classification problems, however the proposed approach can be generalized to multi-class classification problems easily. In future, we want to extend this work to various other domains such as computer vision.

References

1. B. Ganter and R. Wille, *Formal Concept Analysis, Mathematical Foundations*. Berlin,Heidelberg,New York: Springer, 1999.
2. R. Caruana, Y. Lou, J. Gehrke, P. Koch, M. Sturm, and N. Elhadad, "Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission," in *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*,

ser. KDD '15. New York, NY, USA: ACM, 2015, pp. 1721–1730. [Online]. Available: http://doi.acm.org/10.1145/2783258.2788613

3. B. Letham, C. Rudin, T. H. McCormick, and D. Madigan, "An interpretable stroke prediction model using rules and bayesian analysis," in *Proceedings of the 17th AAAI Conference on Late-Breaking Developments in the Field of Artificial Intelligence*, ser. AAAIWS'13-17. AAAI Press, 2013, pp. 65–67. [Online]. Available: http://dl.acm.org/citation.cfm?id=2908286.2908308
4. B. Ustun and C. Rudin, "Supersparse linear integer models for optimized medical scoring systems," *Mach. Learn.*, vol. 102, no. 3, pp. 349–391, Mar. 2016. [Online]. Available: https://doi.org/10.1007/s10994-015-5528-6
5. A. Karpathy and L. Fei-Fei, "Deep visual-semantic alignments for generating image descriptions," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 4, pp. 664–676, Apr. 2017. [Online]. Available: https://doi.org/10.1109/TPAMI.2016.2598339
6. K. Xu, J. L. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhutdinov, R. S. Zemel, and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention," in *Proceedings of the 32Nd International Conference on International Conference on Machine Learning - Volume 37*, ser. ICML'15. JMLR.org, 2015, pp. 2048–2057. [Online]. Available: http://dl.acm.org/citation.cfm?id=3045118.3045336
7. D. Baehrens, T. Schroeter, S. Harmeling, M. Kawanabe, K. Hansen, and K.-R. Müller, "How to explain individual classification decisions," *J. Mach. Learn. Res.*, vol. 11, pp. 1803–1831, Aug. 2010. [Online]. Available: http://dl.acm.org/citation.cfm?id=1756006.1859912
8. M. W. Craven and J. W. Shavlik, "Extracting tree-structured representations of trained networks," in *Proceedings of the 8th International Conference on Neural Information Processing Systems*, ser. NIPS'95. Cambridge, MA, USA: MIT Press, 1995, pp. 24–30. [Online]. Available: http://dl.acm.org/citation.cfm?id=2998828.2998832
9. M. T. Ribeiro, S. Singh, and C. Guestrin, ""why should I trust you?": Explaining the predictions of any classifier," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016*, 2016, pp. 1135–1144. [Online]. Available: https://doi.org/10.1145/2939672.2939778