Big Homework OSDA

Oscar L. Mendoza T.

December 17, 2023

Table of Contents

1 Introduction

In this paper will be introduced implementation of a Lazy FCA classification algorithm base on pattern structures. Proposed algorithm was compared with baseline Lazy FCA algorithm and popular models: Random Forest, CatBoost, Logistic Regression and SVC. For comparison I used three popular datasets:

Titanic Dataset: Titanic Dataset (kaggle.com)

Heart Failure Prediction Dataset: Heart Failure Prediction Dataset (kaggle.com)

Breast Cancer Dataset: Breast Cancer Dataset (kaggle.com)

2 Information about the selected datasets

2.1 Titanic Dataset

The Titanic dataset is a collection of information about the passengers who were aboard the Titanic when it sank on its maiden voyage in 1912. The dataset contains 12 columns, each with a different piece of information about the passengers. Names of the columns and a description are listed below.

- PassengerId: A unique identifier for each passenger

- Survived: Whether or not the passenger survived (0 = did not survive, 1 = survived)

- Pclass: The passenger's class on the ship (1 = first class, 2 = second class, 3 = third class)

- Name: The passenger's name

- Sex: The passenger's gender

- Age: The passenger's age in years (fractional if less than 1)

- SibSp: The number of siblings/spouses the passenger had aboard the ship

- Parch: The number of parents/children the passenger had aboard the ship

- Ticket: The passenger's ticket number

- Fare: The fare the passenger paid for their ticket

- Cabin: The cabin number assigned to the passenger (if any)

- Embarked: The port where the passenger embarked (C = Cherbourg, Q = Queenstown, S = Southampton)

2.2 Heart Failure Prediction Dataset

The Heart Failure Prediction Dataset contains information about patients who were tested for heart disease. The dataset contains 12 columns, each with a different piece of information about the patients. Names of the columns and a description are listed below.

- Age: The patient's age in years

- Sex: The patient's gender (1 = male, 0 = female)

- ChestPainType: The type of chest pain the patient experienced (1 = typical angina, 2 = atypical angina, 3 = non-anginal pain, 4 = asymptomatic)

- RestingBP: The patient's resting blood pressure (in mm Hg)

- Cholesterol: The patient's cholesterol level (in mg/dl)

- FastingBS: Whether or not the patient's fasting blood sugar was greater than 120 mg/dl (1 = true, 0 = false)

- RestingECG: The results of the patient's resting electrocardiogram (0 = normal, 1 = having ST-T wave abnormality, 2 = showing probable or definite left ventricular hypertrophy)

- MaxHR: The patient's maximum heart rate achieved during exercise

- ExerciseAngina: Whether or not the patient experienced angina during exercise (1 = yes, 0 = no)

- Oldpeak: The ST depression induced by exercise relative to rest

- ST_Slope: The slope of the peak exercise ST segment (1 = upsloping, 2 = flat, 3 = downsloping)

- HeartDisease: Whether or not the patient has heart disease (0 = no, 1 = yes)

2.3 Breast Cancer Dataset

The Breast Cancer Dataset contains information about breast cancer tumors. The dataset contains 32 columns, with the first column being an index and the last column being the target variable (y) indicating whether the tumor is malignant or benign.

- Unnamed: 0: The index of the data point

- x.radius_mean: Mean radius of the tumor

- x.texture_mean: Mean texture of the tumor

- x.perimeter_mean: Mean perimeter of the tumor

- x.area_mean: Mean area of the tumor

- x.smoothness_mean: Mean smoothness of the tumor

- x.compactness_mean: Mean compactness of the tumor

- x.concavity_mean: Mean concavity of the tumor

- x.concave_pts_mean: Mean number of concave points of the tumor

- x.symmetry_mean: Mean symmetry of the tumor

- x.fractal_dim_mean: Mean fractal dimension of the tumor

- x.radius_se: Standard error of the radius of the tumor

- x.texture_se: Standard error of the texture of the tumor

- x.perimeter_se: Standard error of the perimeter of the tumor

- x.area_se: Standard error of the area of the tumor

- x.smoothness_se: Standard error of the smoothness of the tumor

- x.compactness_se: Standard error of the compactness of the tumor

- x.concavity_se: Standard error of the concavity of the tumor

- x.concave_pts_se: Standard error of the number of concave points of the tumor

- x.symmetry_se: Standard error of the symmetry of the tumor
- x.fractal_dim_se: Standard error of the fractal dimension of the tumor
- x.radius_worst: Worst radius of the tumor
- x.texture_worst: Worst texture of the tumor
- x.perimeter_worst: Worst perimeter of the tumor
- x.area_worst: Worst area of the tumor
- x.smoothness_worst: Worst smoothness of the tumor
- x.compactness_worst: Worst compactness of the tumor
- x.concavity_worst: Worst concavity of the tumor
- x.concave_pts_worst: Worst number of concave points of the tumor
- x.symmetry_worst: Worst symmetry of the tumor
- x.fractal_dim_worst: Worst fractal dimension of the tumor
- y: The target variable indicating whether the tumor is malignant or benign (0 = benign, 1 = malignant)

3 Classification Using Machine Learning Tools

After selecting 3 datasets I proceed to use Machine learning models. The models under evaluation include decision tree, random forest, XGBoost, k-NN, Naive Bayes, and logistic regression. These models have been selected for their widespread usage and proven performance in classification tasks.

To evaluate the performance of these models, accuracy tests and F1 scores were calculated for each dataset. The accuracy test measures the proportion of correctly classified instances, providing an overall assessment of the model's predictive accuracy. On the other hand, the F1 score considers both precision and recall, providing a balanced measure of the model's performance, particularly in scenarios with imbalanced classes.

By utilizing these evaluation metrics, we gain valuable insights into the strengths and weaknesses of each model when applied to the specific datasets. This information is crucial for making informed decisions about model selection and deployment in practical applications.

The following table(Figure 1) presents a concise summary of the accuracy test results and F1 scores obtained from evaluating the aforementioned machine learning models on the three distinct datasets.

	Random Forest		Decision Tree		k-NN		Naïve Bayes		X G boost		Catboost		Logistic regression	
	Accuracy	f1-score	Accuracy	f1-score	Accuracy	f1-score	Accuracy	f1-score	Accuracy	f1-score	Accuracy	f1-score	Accuracy	f1-score
Titanic Dataset	1.0	1.0	1.0	1.0	0.78	0.85	1.0	1.0	1.0	1.0	1.00	1.00	1.0	1.0
Breast Cancer Dataset	0.99	0.99	0.98	0.98	1.0	0.99	0.95	0.96	1.0	1.0	1.00	1.00	0.98	0.98
Hearth desease diagnostic Dataset	0.64	0.70	0.68	0.66	0.66	0.69	0.66	0.64	0.65	0.72	0.65	0.71	0.68	0.67

Figure 1 – Results of Accuracy and F1 score using different ML tools

4 Lazy-FCA classification with binary attributes

For the use Lazy classification I binarized every datasets and after binarization of the datasets I performed the Lazy Classfication Algorithm 9 times with every dataset. I performed 9 times this algorithm because I performed 3 trials with different alpha values(0, 0.5 and 1) every method of the algorithm(standard, standard-support and ratio-support) and as a result we can see the values of the accuracy test and f1 score in the Figure 2.

Dataset	Methods	Scores	Alpha Values		
			0	0,5	1
TITANIC DATASET	Standard	Accuracy Score	0,98	0,98	0
		F1 score	0,97	0,97	0
	Standard-support	Accuracy Score	0	0	0
		F1 score	0	0	0
	Ratio-support	Accuracy Score	0,67	0,67	0,67
		F1 score	0	0	0
Heart Disease Prediction Dataset	Standard	Accuracy	0,66	0,66	0
		F1 score	0,73	0,73	0
	Standard-support	Accuracy Score	0	0	0,28
		F1 score	0	0	0,33
	Ratio-support	Accuracy Score	0,31	0,31	0,31
		F1 score	0	0	0
Breast Cancer Dataset	Standard	Accuracy Score	0,88	0,88	0
		F1 score	0,81	0,81	0
	Standard-support	Accuracy Score	0	0	0,12
		F1 score	0	0	0,17
	Ratio-support	Accuracy Score	0,12	0,12	0,12
		F1 score	0,17	0,17	0,17

Figure 2 – Accuracy and F1 score of the Lazy Classification Algorithm with Binary Structures

5 Lazy-FCA classification with pattern structures

In the same way as Lazy Classification with binary structures I performed 9 times this algorithm because I performed 3 trials with different alpha values(0, 0.5 and 1) every method of the algorithm(standard, standard-support and ratio-support) and as a result we can see the values of the accuracy test and f1 score in the Figure 6.

Dataset	Methods	Scores	Alpha Values		
			0	0,5	1
TITANIC DATASET	Standard	Accuracy Score	0.98	0.98	0
		F1 score	0.97	0.97	0
	Standard-support	Accuracy Score	0.2	0	0
		F1 score	0.1	0	0
	Ratio-support	Accuracy Score	0.67	0.67	0.67
		F1 score	0	0	0
Heart Disease Prediction Dataset	Standard	Accuracy	0.66	0.66	0
		F1 score	0.73	0.73	0
	Standard-support	Accuracy Score	0.33	0.33	0.33
		F1 score	0.28	0.28	0.28
	Ratio-support	Accuracy Score	0.31	0	0
		F1 score	0.21	0	0
Breast Cancer Dataset	Standard	Accuracy Score	0.87	0.87	0
		F1 score	0.81	0.81	0
	Standard-support	Accuracy Score	0	0	0.12
		F1 score	0	0	0.12
	Ratio-support	Accuracy Score	0	0	0.68
		F1 score	0	0	0.67

Figure 3 – Accuracy and F1 score of the Lazy Classification Algorithm with Pattern Structures

6 Conclusion

In conclusion after a detailed comparison with the current configuration of parameters of the algorithms we can see that ML models like decision tree, random forest, XGBoost, k-NN, Naive Bayes, and logistic regression are more accurate in the moment of predictions when we are working with de Titanic Dataset, obtaining really similar results of accuracy and f1 score. In the case of the Breast Cancer Dataset we can see a clear difference but not very big, by using the different ML models of the first part and the Lazy Classification Algorithm. And using Breast Cancer Dataset there are no difference between the results of the different ML models and the Lazy Classification algorithm. Finally, it is worth emphasizing that all three datasets exhibit no noteworthy variations when evaluated based on f1 and accuracy scores, particularly when the Lazy Classification Algorithm is employed for analyzing binary structures and patterns.