FCA project report. Ordered sets in data analysis

Student: Atymkhanova Madina

Link: `https://github.com/madina9997/FCA_project`

The aim of the project

1. For chosen dataset implement lazy-formal concept analysis and for test dataset solve classification problem
2. Varying hyperparameters of built model analize what values fit it better.

The dataset

`https://archive.ics.uci.edu/ml/datasets/congressional+voting+records`

Data Set Characteristics:	Multivariate	**Number of Instances:**	435	**Area:**	Social
Attribute Characteristics:	Categorical	**Number of Attributes:**	16	**Date Donated**	1987-04-27

This data set includes votes for each of the U.S. House of Representatives Congressmen on the 16 key votes identified by the CQA. The CQA lists nine different types of votes: voted for, paired for, and announced for (these three simplified to yea), voted against, paired against, and announced against (these three simplified to nay), voted present, voted present to avoid conflict of interest, and did not vote or otherwise make a position known (these three simplified to an unknown disposition).

Values encoding:

- “y” - stands for ”yea”(the meaning denoted in the above)
- “n” - stands for ”nay”
- “?” - stands for ”not yea nor nay”

Additional information:

- Number of Instances: 435 (267 democrats, 168 republicans)
- Number of Attributes: 16 + class name = 17 (all Boolean valued)
- Attribute Information:

    1. Class Name: 2 (democrat, republican)
    2. handicapped-infants: 2 (y,n)
    3. water-project-cost-sharing: 2 (y,n)
    4. adoption-of-the-budget-resolution: 2 (y,n)
    5. physician-fee-freeze: 2 (y,n)
    6. el-salvador-aid: 2 (y,n)
    7. religious-groups-in-schools: 2 (y,n)
    8. anti-satellite-test-ban: 2 (y,n)
    9. aid-to-nicaraguan-contras: 2 (y,n)
    10. mx-missile: 2 (y,n)
    11. immigration: 2 (y,n)
    12. synfuels-corporation-cutback: 2 (y,n)
    13. education-spending: 2 (y,n)
    14. superfund-right-to-sue: 2 (y,n)
    15. crime: 2 (y,n)
    16. duty-free-exports: 2 (y,n)
    17. export-administration-act-south-africa: 2 (y,n)

Lazy-FCA approach

In lazy-FCA approach was implemented simple function for decision making which performed with the rusults in below.

$$Positive\ support = \frac{1}{|G^+|} \sum_{i=0}^{|G^+|} |g' \cap g_i^+|$$

$$Negative\ support = \frac{1}{|G^-|} \sum_{i=0}^{|G^-|} |g' \cap g_i^-|$$

Where for positive context(G^+) and negative context(G^-) stand elements(voters) which belongs for 'republican' party and 'democratic' party respectively.

The intersection between test element for classifying(g') and element from context (g_i^+ / g_i^-) performed as comparing for an equality values attribute-wise. In case of '?' value of attribute for resulting pattern for this pair of elements intersection would contain in respective attribute pair of two values.

When searching for pattern inclusion in opposite context we also need to compare "votes"attribute-wise. If the pattern is contained in some element in opposite contex we will not then consider it as a pattern and it doesnt have influence on support.

Comparing "positive support"and "negative support"we would take a decision in preference of that context for which support is bigger.

Hyperparameter inclusion

There is a way to modify that algorithm on a stage when supports are compared. The conditions changes from simple inequalities between support:

```
...
if pos_supp > neg_supp:
    pos_classified += 1
```

```
elif pos_supp < neg_supp:
    neg_classified += 1
else:
    undef_class += 1
...
```

to :

```
...
if pos_supp > hyperparam1 + neg_supp:
    ...
elif pos_supp < hyperparam2 + neg_supp:
    ...
```

The hyperparameters corresponse with the similarity allowance between parties in advance.

Varying set of pairs of hypermarameters we obtain the best result (for all partitions of initial dataset with respect to cross-validation procedure) which is:

```
hyperparam1 = 12  hyperparam2 = -9
precision = 1.0
recall = 1.0
accuracy = 1.0
```

Cross-validation

Before starting lazy-fca dataset is been separated on some number of parts (which is a parameter, on the outer cycle is varied). Consequentially test set is chosen among parts. Rest parts become training set.

The result appeared to be quite robust for changing number of parts.

Results

Here provided few result on described dataset:

```
hyperparam1 = -4  hyperparam2 = 13
num_part_crossval = 18
precision = 0.8299374236874236
recall = 0.9802943969610636
accuracy = 0.9000000000000001
```

```
hyperparam1 = 10  hyperparam2 = -5
num_part_crossval = 25
precision = 0.9816666666666667
recall = 0.9736666666666668
accuracy = 0.9850606060606062
```

```
hyperparam1 = 0  hyperparam2 = 0
num_part_crossval = 4
precision = 0.8844724968347484
recall = 0.9471996983447636
accuracy = 0.928644240570846
```

State-of-the-art algoritms

Likelihood of the obtained results was estimated in a comparison with the state-of-the-art classifiers such as:

- Random Forest Classifier
- Logistic Regression

Following results were obtained:

- For Random Forest Classifier:

```
random forest accuracy with crossval
num_part_crossval = 7
accuracy = 0.9682539682539683
accuracy = 0.9682539682539683
```

```
accuracy = 0.9682539682539683
accuracy = 0.9682539682539683
accuracy = 0.9841269841269841
accuracy = 0.9206349206349206
accuracy = 0.9298245614035088
```

- Logistic Regression

```
linreg accuracy with crossval
num_part_crossval = 7
accuracy = 0.9682539682539683
accuracy = 0.9365079365079365
accuracy = 0.9365079365079365
accuracy = 0.9682539682539683
accuracy = 0.9523809523809523
accuracy = 0.9365079365079365
accuracy = 0.8947368421052632
```

So we may conclude in this paragraph that obtained with lazy-FCA result and result obtained with SOTA methods are similar, what points out on a reliability of implemented algorithm.

On a picture below element of plotting the distribution of votes among democrats and republicans for a concrete law (the numbers above subplots).

Looking at a whole picture of subplots there is in general big difference between amount of democrates who supported adoption of respective law (they listed in a second page) and republicans.

For example the 5th attribute which stands for ”el-salvador-aid”law illustrates how clearly parties separated - most of democrats voted against whereas republican supported the law.

From this observations we may conclude that the good occuracy of the model is partially might be provided just by good distinction between decisions parties make.

It provides good metric results for classification models.