|  |  |
| --- | --- |
| lu135925on3bu_tmp_3360867a00ce4d37 | **Министерство науки и высшего образования Российской Федерации Федеральное государственное бюджетное образовательное учреждение высшего образования** **«Московский государственный технический университет имени Н.Э. Баумана** **(национальный исследовательский университет)» (МГТУ им. Н.Э. Баумана)** |

ФАКУЛЬТЕТ Информатика и системы управления и искусственный интеллект

КАФЕДРА                  Системы обработки информации и управления

**Домашнее задание №1**

**По курсу**

**«Оптимизация баз данных систем машинного обучения»**

**«Обнаружение функциональных зависимостей»**

**Вариант 4**

Выполнил: Журавлев Н.В.

Группа: ИУ5-14М

Дата: 29.09.2023

Проверил:

Плужникова О. Ю.

2023 г.

# **Задание**

Для каждого заданного набора данных:

1. Определите функциональные зависимости с помощью точного гибридного алгоритма HyFD (алгоритм может выполняться некоторое время);
2. Определите функциональные зависимости с помощью приближённого алгоритма AIDFD в течение 1 секунды его работы;
3. После выполнения алгоритма AIDFD постройте график зависимости прироста числа элементов отрицательного покрытия от номера итерации *k* на основе данных, отображаемых в окне "Adding Context for backend";
4. Рассчитайте показатель полноты (П) на основе полученных точных (HyFD) и приближённых (AIDFD) функциональных зависимостей.

# **Набор данных по варианту**

|  |  |  |  |
| --- | --- | --- | --- |
| ИУ5-14М 2023 | | Варианты наборов данных | |
| 4 | Журавлев Николай Вадимович | 4  (07 adult.zip) | 20  (24 real+estate+valuation+data+set.zip) |

# **Описание наборов данных**

**adolt.csv** – набор данных из базы данных переписи 1994 года, который был выполнен Барри Беккером, чтобы спрогнозировать, превысит ли доход 50 тысяч долларов в год на основе данных переписи населения. Набор содержит:

1. age – возраст; значения - любые численные
2. workclass – доход; значения - Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, Without-pay, Never-worked
3. fnlwgt - примерная оценка количества людей, которое представляет каждая строка данных; значения - любые численные
4. education – уровень образования; значения - Bachelors, Some-college, 11th, HS-grad, Prof-school, Assoc-acdm, Assoc-voc, 9th, 7th-8th, 12th, Masters, 1st-4th, 10th, Doctorate, 5th-6th, Preschool
5. education-num – длительность обучения; значения - любые численные
6. marital-status – семейное положение; значения - Married-civ-spouse, Divorced, Never-married, Separated, Widowed, Married-spouse-absent, Married-AF-spouse
7. occupation – род деятельности; значения - Tech-support, Craft-repair, Other-service, Sales, Exec-managerial, Prof-specialty, Handlers-cleaners, Machine-op-inspct, Adm-clerical, Farming-fishing, Transport-moving, Priv-house-serv, Protective-serv, Armed-Forces
8. relationship - отношения; значения - Wife, Own-child, Husband, Not-in-family, Other-relative, Unmarried
9. race - раса; значения - White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black
10. sex - пол; значения - Female, Male
11. capital-gain – прирост капитала; значения - любые численные
12. capital-loss – потеря капитала; значения - любые численные
13. hours-per-week – количество рабочих часов в неделю; значения - любые численные
14. native-country – родная страна; значения - United-States, Cambodia, England, Puerto-Rico, Canada, Germany, Outlying-US(Guam-USVI-etc), India, Japan, Greece, South, China, Cuba, Iran, Honduras, Philippines, Italy, Poland, Jamaica, Vietnam, Mexico, Portugal, Ireland, France, Dominican-Republic, Laos, Ecuador, Taiwan, Haiti, Columbia, Hungary, Guatemala, Nicaragua, Scotland, Thailand, Yugoslavia, El-Salvador, Trinadad&Tobago, Peru, Hong, Holand-Netherlands

**Estate\_valuation.csv** – набор рыночных данных по оценке недвижимости, собранный в Синдианском округе, город Нью-Тайбэй, Тайвань. Набор содержит:

1. X1 transaction date – дата заключения сделки; значения - дата в формате год, прошедшая часть (например, 2013,250 = март 2013 г., 2013,500 = июнь 2013)
2. X2 house age – возраст дома; значения – любые численные
3. X3 distance to the nearest MRT – метров до ближайшей станции метро; значения - любые численные
4. X4 number of convenience stores - количество магазинов повседневного спроса в шаговой доступности; значения - любые численные
5. X5 latitude – широта, на которой расположен дом; значения - любые численные
6. X6 longitude – долгота, на которой расположен дом; значения - любые численные
7. Y house price of unit area - стоимость дома за единицу площади; значения - любые численные

# **Ход работы**

### Для adolt.csv

Данные изначально находились в архиве с расширением data и с файлом описанием колонок в формате name.

### Очистка данных

Датасет был преобразован в csv файл. Затем с помощью программы excel было удалено 24 дубликата. Файл заголовки не содержал. Колонка ключи в данном датасете отсевает.

### Работа в Metanome

Файлы, содержащие реализацию алгоритмов HyFD и AIDFD были перемещены по пути …\backend\WEB-INF\classes\algorithms.

Затем файла с данными был перемещён в директорию по пути: …\backend\WEB-INF\classes\inputData.

Так как в файле не имеются заголовки, то необходимо при выборе датасета в программе убрать галочку “Has Header”.

### Число ФЗ, полученных алгоритмами

Результаты использования для алгоритма HyFD, представлена на рис.1.

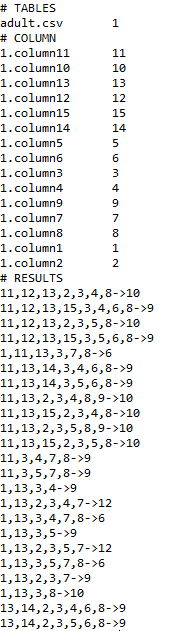


Рисунок 1

Всего найдено 78 функциональных зависимостей.

Результаты использования для алгоритма AIDFD представлены на рис. 2.

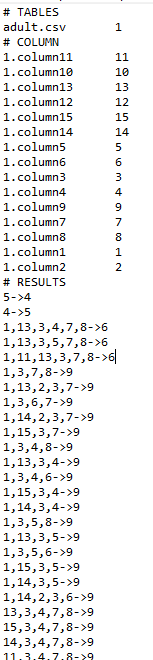


Рисунок 2

Всего найдено 78 функциональные зависимости.

**Вывод:** Результат работы алгоритмов совпал.

### График зависимости прироста числа элементов отрицательного покрытия от номера итерации k

Данные, получившиеся в результате работы алгоритма представлены в таблице ниже:

|  |  |  |
| --- | --- | --- |
| **k** | **negCoverSize** | **negCoverRatio** |
| 1 | 6528 | 1.7976931348623157E308 |
| 2 | 7345 | 0.1251531862745098 |
| 3 | 7733 | 0.0528205105513955 |
| 4 | 7965 | 0.03000129315187897 |
| 5 | 8142 | 0.02222222222222223 |
| 6 | 8254 | 0.013755833947433063 |
| 7 | 8329 | 0.009086503513448025 |

График зависимости прироста числа элементов отрицательного покры-тия от номера итерации:

### Расчет показателя полноты для приближенных ФЗ, полученных за 1 секунду работы алгоритма AIDFD

Показатель полноты рассчитывается по следующей формуле:

,

где - всё множество нетривиальных минимальных функциональных зависимостей (будем называть их истинными), - множество приближённых функциональных зависимостей, возвращаемых алгоритмом, |*X*| обозначает число элементов в множестве X.

В данном случае:

### Для estate\_valuation.csv

Данные изначально находились в архиве с расширением xlsx.

### Очистка данных

Файл был преобразован в csv. Дубликатов датасет не имеет. Присутствовал столбец с идентификатором записей, который затем удалён.

### Работа в Metanome

Так же файла с данными был перемещён в директорию по пути: …\backend\WEB-INF\classes\inputData.

Так как в файле имеются заголовки, то необходимо при выборе датасета в программе вернуть галочку “Has Header”.

### Число ФЗ, полученных алгоритмами

Результаты использования для алгоритма HyFD, представлена на рис.3.

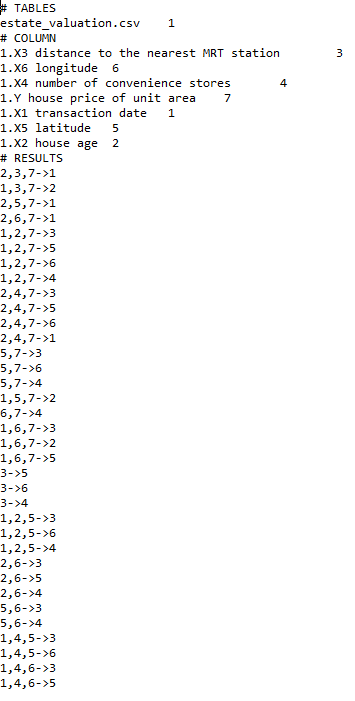


Рисунок 3

Всего найдено 35 функциональных зависимостей.

Результаты использования для алгоритма AIDFD, представлена на рис.4.

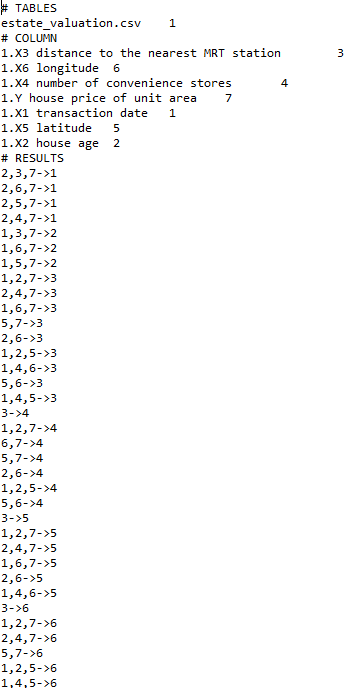


Рисунок 4

Всего найдено 35 функциональных зависимостей.

**Вывод:** Результат работы алгоритмов совпал.

### График зависимости прироста числа элементов отрицательного покрытия от номера итерации k для алгоритма AIDFD

Данные, получившиеся в результате работы алгоритма представлены в таблице ниже:

|  |  |  |
| --- | --- | --- |
| **k** | **negCoverSize** | **negCoverRatio** |
| 1 | 25 | 1.7976931348623157E308 |
| 2 | 26 | 0 |
| 3 | 26 | 0 |
| 4 | 26 | 0 |
| 5 | 26 | 0 |
| … | … | … |
| 67 | 26 | 0 |

График зависимости прироста числа элементов отрицательного покрытия от номера итерации:

### Расчет показателя полноты для приближенных ФЗ, полученных за 1 секунду работы алгоритма AIDFD

По формуле, представленной при описания предыдущего датасета, рассчитываем показатель полноты: