**EvayAIDataMining**

**接口使用说明**

山东亿云信息技术有限公司

创新研究院

2019年4月

目录

[一、概述 3](#_Toc10013532)

[二、接口模块 3](#_Toc10013533)

[三、接口说明 3](#_Toc10013534)

[1、 √数据预处理 3](#_Toc10013535)

[2、 √机器学习算法库 13](#_Toc10013536)

# 一、概述

EvayAIDataMining

# 二、接口模块

1、数据预处理

2、机器学习算法库

3、模型应用

4、模型评估

5、文本分析

文本分类

文本聚类

文本分析——模型应用

6、模型库

招投标分类模型

百科人物信息抽取模型

# 三、接口说明

## √数据预处理

**请求方法**

Post:

<http://10.20.15.51:8080/preprocess>

{

"input\_sql\_url":{

"sql\_type":"mysql",

"username":"root",

"password":"BigData@2018",

"host":"10.20.5.49:3306",

"database":"data\_mining\_DB"

},

"output\_sql\_url":{

"sql\_type":"mysql",

"username":"root",

"password":"123456",

"host":"127.0.0.1:3306",

"database":"data\_mining\_DB"

},

"temp\_table": "ml\_info\_item",

"new\_table": "ml\_info\_type\_copy\_data\_invalid110",

"process": ["data\_invalid","contain"],

"params": ["OPEN\_CONDITION", "equal", "无"]

}

**参数说明**

|  |  |  |
| --- | --- | --- |
| **Json** | **含义** | **说明** |
| input\_sql\_url | 输入数据库连接信息 | "input\_sql\_url":{  "sql\_type":"mysql",  "username":"root",  "password":"BigData@2018",  "host":"10.20.5.49:3306",  "database":"data\_mining\_DB" },  sql\_type支持mysql或oscar |
| output\_sql\_url | 输出数据库连接信息 | "output\_sql\_url":{  "sql\_type":"mysql",  "username":"root",  "password":"BigData@2018",  "host":"10.20.5.49:3306",  "database":"data\_mining\_DB" },  sql\_type支持mysql |
| temp\_table | 数据表 | 表名 |
| new\_table | 结果表 | 表名 |
| process | 预处理方法：  [功能，处理方法] | **空值：**注list顺序不可变  (0) ["data\_null","donothing"] 不做处理  (1) ["data\_null","save"] 保存空值  (2) ["data\_null","drop"] 删除空值  (3) ["data\_null","mean\_data"] 填充均值(数值)  (4) ["data\_null","median\_data"] 填充中位数(数值)  (5) ["data\_null","mode\_data"] 填充众数(数值)  (6) ["data\_null","constant"] 填充固定值  **重复值：**  (1) ["data\_duplicate","save"] 保存重复值  (2) ["data\_duplicate ","drop"] 删除重复值  **无效数据：**  (1) ["data\_invalid","contain"] 包含处理  (2) ["data\_invalid","length"] 长度处理  (3) ["data\_invalid","number"] 数值处理  **基于正则数据过滤：**  (1) ["re\_data\_filter","id\_check"] 身份证校验  (2) ["re\_data\_filter","email\_check"] 邮箱校验  (3) ["re\_data\_filter","telephone\_check"] 电话号码校验  (4) ["re\_data\_filter","regex\_check"] 输入正则表达式校验  **异常值：**  (0) ["data\_abnormal","drop"] 删除异常值  (1) ["data\_abnormal","mean\_data"] 异常值视为均值  (2) ["data\_abnormal","miss\_data"] 异常值视为缺失值  (3) ["data\_abnormal","donothing"] 不做处理 |
| params | 参数：[列名,处理方法] | **data\_null：**  (0-5)：["列名"]，["A","B"]，支持一列或多列  (6)：["列名","固定值"]，["A","B","10"]，支持一列或多列，最后为固定值，注list顺序不可变  **data\_duplicate：**  ["列名"]，["A","B"]，支持一列或多列  **data\_invalid：**  ["列名","方法","字符串/数值"] 注list顺序不可变  (1)字符串包含过滤：例：["A","in","路线明细表"]，仅支持一列，支持：包含 in，不包含：notin，等于：equal  (2)长度过滤：例：["A",">","10"]，仅支持一列，支持>、<、>=、<=、=、!=，字符长度，注list顺序不可变  (3)数值过滤：例：["A",">","10"]，仅支持一列，支持>、<、>=、<=、=、!=，数值大小，注list顺序不可变  **re\_data\_filter：**  ["列名","方法"] 注list顺序不可变，用于身份证、邮箱、电话号码校验传参 或  ["列名","方法","字符串"] 注list顺序不可变，用于正则表达式校验  (1)身份证校验：["A","check\_true"]，仅支持一列，方法支持check\_false、check\_true  (2)邮箱校验：["列名","check\_false"]，仅支持一列，方法支持check\_false、check\_true  (3)电话号码校验：["列名","check\_false"]，仅支持一列，方法支持check\_false、check\_true  (4)输入正则表达式校验：["email","check\_false","163.com"]，仅支持一列，支持check\_false、check\_true，支持输入正则表达式(字符串)  **data\_abnormal：**  ["列名","列名","zscore"] 注list顺序不可变，支持一列或多列  (0-3)：["列名","zscore"]，["A","B","mean\_data"]，支持一列或多列，最后为zscore值，注list顺序不可变 |

**返回值**

在接收到客户端发送的调用请求后，返回以json的方式返回处理结果。

**创建成功示例：**

返回数据，创建新表

{  
 "Data": "",  
 "Mesg": "write to ml\_info\_type\_copy\_data\_invalid102 success",  
 "Code": 0  
}

返回空数据，不创建新表

{  
 "Data": "",  
 "Code": -1,  
 "Mesg": "Null data is None"  
}

**创建失败示例：**

{  
 "Data": "",  
 "Code": 1,  
 "Mesg": "(pymysql.err.ProgrammingError) (1146, \"Table 'data\_mining\_db.ml\_info\_type\_copy' doesn't exist\")\n[SQL: select \* from ml\_info\_type\_copy]\n(Background on this error at: http://sqlalche.me/e/f405)"  
}

**测试用例**

1. 缺失值处理接口

(1) ml\_info\_type表中table\_name为空的，将table\_name字段为空值的全部数据保存出来。

Post:

<http://10.20.15.51:8080/preprocess>

{

"input\_sql\_url":{

"sql\_type":"mysql",

"username":"root",

"password":"BigData@2018",

"host":"10.20.5.49:3306",

"database":"data\_mining\_DB"

},

"output\_sql\_url":{

"sql\_type":"mysql",

"username":"root",

"password":"123456",

"host":"127.0.0.1:3306",

"database":"data\_mining\_DB"

},

"temp\_table": "ml\_info\_type\_copy",

"new\_table": "ml\_info\_type\_copy\_data\_invalid101",

"process": ["data\_null", "save"],

"params": ["TABLE\_NAME"]

}

(2) ml\_info\_type表中table\_name为空的，将table\_name字段为空值的填补“空值”。  
Post:

<http://10.20.15.51:8080/preprocess>

{

"input\_sql\_url":{

"sql\_type":"mysql",

"username":"root",

"password":"BigData@2018",

"host":"10.20.5.49:3306",

"database":"data\_mining\_DB"

},

"output\_sql\_url":{

"sql\_type":"mysql",

"username":"root",

"password":"123456",

"host":"127.0.0.1:3306",

"database":"data\_mining\_DB"

},

"temp\_table": "ml\_info\_type\_copy",

"new\_table": "ml\_info\_type\_copy\_data\_invalid101",

"process": ["data\_null","constant"],

"params": ["TABLE\_NAME", " "]

}

2. 数据去重接口

(1) ml\_info\_type表中provider重复的数据展示出来  
Post:

<http://10.20.15.51:8080/preprocess>

{

"input\_sql\_url":{

"sql\_type":"mysql",

"username":"root",

"password":"BigData@2018",

"host":"10.20.5.49:3306",

"database":"data\_mining\_DB"

},

"output\_sql\_url":{

"sql\_type":"mysql",

"username":"root",

"password":"123456",

"host":"127.0.0.1:3306",

"database":"data\_mining\_DB"

},

"temp\_table": "ml\_info\_type\_copy",

"new\_table": "ml\_info\_type\_copy\_data\_invalid101",

"process": ["data\_duplicate","save"],

"params": ["PROVIDER"]

}

3. 数据过滤接口

(1) ml\_info\_type表中type\_desc字符数小于10的。  
Post:

<http://10.20.15.51:8080/preprocess>

{

"input\_sql\_url":{

"sql\_type":"mysql",

"username":"root",

"password":"BigData@2018",

"host":"10.20.5.49:3306",

"database":"data\_mining\_DB"

},

"output\_sql\_url":{

"sql\_type":"mysql",

"username":"root",

"password":"123456",

"host":"127.0.0.1:3306",

"database":"data\_mining\_DB"

},

"temp\_table": "ml\_info\_type\_copy",

"new\_table": "ml\_info\_type\_copy\_data\_invalid101",

"process": ["data\_invalid","length"],

"params": ["TYPE\_DESC", "<", "10"]

}

(2) ml\_info\_item表里open\_condition字段为空字符串的全部数据保存出来。  
Post:

<http://10.20.15.51:8080/preprocess>

{

"input\_sql\_type":"mysql",

"input\_sql\_url":{

"username":"root",

"password":"123456",

"host":"127.0.0.1:3306",

"database":"data\_mining\_DB"

},

"output\_sql\_type":"mysql",

"output\_sql\_url":{

"username":"root",

"password":"BigData@2018",

"host":"10.20.5.49:3306",

"database":"data\_mining\_DB"

},

"temp\_table": "ml\_info\_item",

"new\_table": "ml\_info\_type\_copy\_data\_invalid111",

"process": ["data\_invalid","contain"],

"params": ["OPEN\_CONDITION", "equal", "无"]

}

4. 基于正则的数据过滤

（1）身份证校验

Post:

<http://10.20.15.51:8080/preprocess>

{

"input\_sql\_url":{

"sql\_type":"mysql",

"username":"root",

"password":"123456",

"host":"127.0.0.1:3306",

"database":"data\_mining\_DB"

},

"output\_sql\_url":{

"sql\_type":"mysql",

"username":"root",

"password":"123456",

"host":"127.0.0.1:3306",

"database":"data\_mining\_DB"

},

"temp\_table": "rc\_jbxx\_20190301",

"new\_table": "rc\_re\_data\_filter\_phone0",

"process": ["re\_data\_filter","id\_check"],

"params": ["identification\_id","check\_false"]

}

（2）邮箱校验

Post:

<http://10.20.15.51:8080/preprocess>

{

"input\_sql\_url":{

"sql\_type":"mysql",

"username":"root",

"password":"123456",

"host":"127.0.0.1:3306",

"database":"data\_mining\_DB"

},

"output\_sql\_url":{

"sql\_type":"mysql",

"username":"root",

"password":"123456",

"host":"127.0.0.1:3306",

"database":"data\_mining\_DB"

},

"temp\_table": "rc\_jbxx\_20190301",

"new\_table": "rc\_re\_data\_filter\_email0",

"process": ["re\_data\_filter","email\_check"],

"params": ["email","check\_false"]

}

（3）电话号码校验

Post:

<http://10.20.15.51:8080/preprocess>

{

"input\_sql\_url":{

"sql\_type":"mysql",

"username":"root",

"password":"123456",

"host":"127.0.0.1:3306",

"database":"data\_mining\_DB"

},

"output\_sql\_url":{

"sql\_type":"mysql",

"username":"root",

"password":"123456",

"host":"127.0.0.1:3306",

"database":"data\_mining\_DB"

},

"temp\_table": "rc\_jbxx\_20190301",

"new\_table": "rc\_re\_data\_filter\_phone0",

"process": ["re\_data\_filter"," telephone\_check"],

"params": [mobile\_phone","check\_false"]

}

（4）输入正则表达式校验

Post:

<http://10.20.15.51:8080/preprocess>

{

"input\_sql\_url":{

"sql\_type":"mysql",

"username":"root",

"password":"123456",

"host":"127.0.0.1:3306",

"database":"data\_mining\_DB"

},

"output\_sql\_url":{

"sql\_type":"mysql",

"username":"root",

"password":"123456",

"host":"127.0.0.1:3306",

"database":"data\_mining\_DB"

},

"temp\_table": "rc\_jbxx\_20190301",

"new\_table": "rc\_re\_data\_filter\_email2",

"process": ["re\_data\_filter","regex\_check"],

"params": ["email","check\_true","163.com"]

}

5. 异常值处理

Post:

<http://10.20.15.51:8080/preprocess>

{

"input\_sql\_url":{

"sql\_type":"mysql",

"username":"root",

"password":"123456",

"host":"127.0.0.1:3306",

"database":"data\_mining\_DB"

},

"output\_sql\_url":{

"sql\_type":"mysql",

"username":"root",

"password":"123456",

"host":"127.0.0.1:3306",

"database":"data\_mining\_DB"

},

"temp\_table": "data\_statistics",

"new\_table": "ds\_data\_abnormal\_id1",

"process": ["data\_abnormal","mean\_data"],

"params": ["var1","var3","1"]

}