Hans Journal of Data Mining 数据挖掘, 2019, 9(3), 104-108
Published Online July 2019 in Hans. http://www.hanspub.org/journal/hjdm
https://doi.org/10.12677/hjdm.2019.93013

Hans汉斯

# Study on Large Data of Major Elements of Mudstone in Songliao Basin

**Yu Tian**

Key Laboratory for Reservoir Formation of Dense Oil and Shale Oil in Heilongjiang Province, Exploration and Development Research Institute of Daqing Oilfield Co., Ltd., Daqing Heilongjiang
Email: tianyu@petrochina.com.cn

Received: Jun. 26th, 2019; accepted: Jul. 9th, 2019; published: Jul. 16th, 2019

## Abstract

**Mudstone is a common sedimentary rock, which has important value in oil and gas exploration. Firstly, it can be used as a caprock to preserve the fluid in sedimentary basins. Secondly, it can also accumulate a large number of plankton and organic colloids in a certain anoxic environment, thus transforming them into oil-generating reservoirs. Based on the principal element data of 4766 rock samples in Songliao Basin, RapidMiner, a large data analysis software, is used to process and analyze them rapidly. The classification of mudstones in Songliao Basin and the sedimentary conditions of the basin are revealed.**

## Keywords

**Songliao Basin, Mudstone, Major Elements, Large Data, RapidMiner**

# 松辽盆地泥岩主量元素大数据研究

**田　雨**

大庆油田有限责任公司勘探开发研究院，黑龙江省致密油泥页岩油成藏研究重点实验室，黑龙江　大庆
Email: tianyu@petrochina.com.cn

收稿日期：2019年6月26日；录用日期：2019年7月9日；发布日期：2019年7月16日

## 摘　要

**泥岩是一种常见的沉积岩，它在油气勘探中有着重要的价值。首先它可以作为盖层来保留沉积盆地中的流体，其次它也能在一定的乏氧环境中堆积大量的浮游生物及有机胶体从而转化为生油层。本文以松辽盆地4766块岩石样品的主量元素数据为基础，采用大数据分析软件RapidMiner对其进行快速的处理和**

文章引用：田雨. 松辽盆地泥岩主量元素大数据研究[J]. 数据挖掘, 2019, 9(3): 104-108.
DOI: 10.12677/hjdm.2019.93013

**分析，揭示了松辽盆地泥岩的分类及其所反映的盆地沉积条件。**

## 关键词

**松辽盆地，泥岩，主量元素，大数据，RapidMiner**



## 1. 引言

从 20 世纪 60 年代以来，数学地质开始迅速的发展。它是地质学与数学和计算机科学相互渗透、紧密结合而逐步形成的一门地质学的边缘科学[1]。随着全球地质勘探的积累以及分析仪器的发展，海量地质数据迎面而来。鉴于大数据是大容量 、多样性、高速增长、低价值密度的数据集合，传统手段已难于管理和有效发挥其价值[2]，如何快速的处理这些数据并得到其在地质上所表示的一般规律成了科研工作者的难题。

松辽盆地是中国东北部的一个大型中、新生代沉积盆地，地跨黑龙江省、吉林省、辽宁省和内蒙古自治区，是当今世界上最大的典型陆相沉积盆地之一，也是我国最主要的含油气盆地。泥岩作为一种典型的沉积岩，遍布于松辽盆地。而它在地质上也有很重要的存在意义，一是可以作为盖层来保存流体[3]，二是在一定的还原环境中也可以堆积大量的浮游生物等有机体从而转变为生油层。泥岩是沉积岩中数量最多分布最广的岩石，大多数泥岩是在静水环境中沉积的，其颜色和成分常能反映沉积时的介质条件[4]。本文借助 RapidMiner 软件来对松辽盆地泥岩的主量元素数据进行快速的处理和分析。

## 2. RapidMiner 简介

RapidMiner 是世界领先的数据挖掘解决方案，数据挖掘过 程简单，强大和直观。它提供数据挖掘和机器学习程序，其中包括数据加载和转换，数据的预处理和可视化，预测分析和统计建模，评估和部署。它是用 Java 编程语言的。而且提供图形用户界面，用户可以不用编程，通过简单拖拽算子来设计和执行工作流程分析，易于学习和掌[5]。其解决方案覆盖了各个领域，包括汽车、银行、保险、生命科学、制造业、石油和天然气、零售业及快消行业、通讯业、以及公用事业等各个行业。2015 年在 KDnuggets 第 16 届年度数据挖掘大会分析软件投票位中位居第 2，仅次于 R 语言。因为其具备 GUI 特性，所以很适合于数据挖掘的初学者入门。

RapidMiner 中的功能均是通过连接各类算子(operator)形成流程(process)来实现的，整个流程可以看做是工厂车间的生产线，输入原始数据，建造模型，输出结果。其建模的一般流程是：新建一个库(Repository)，选择需要的算子(operator)放入主流程(mainprocess)中，设置算子的相关参数(parameter)，进行算子连接，最后执行流程以得到结果。

## 3. 松辽盆地泥岩数据处理和分析

### 3.1. 数据准备

在地质勘探中，常会使用 X 荧光光谱仪对岩石样品进行元素的定性定量分析，然后利用这些数据大

概的判断一下地质背景、地质构造成因、成岩成矿推断等各种应用。大庆油田在数十年来勘探松辽盆地中积累了大量的元素分析数据，现在我们想要利用 RapidMiner 对其进行一个简单的统计和分类。首先，我们将 4766 块岩石样品的无机元素常量数据导入到系统里面，见图 1。

ExampleSet (4766 examples, 0 special attributes, 16 regular attributes) Filter (4766/4766 examples): all

| Row No. | 井号 | 井深 | 层位 | 岩性 | $Fe_2O_3$ | Mn | Ti | CaO | $K_2O$ | S | P | $SiO_2$ | $Al_2O_3$ | MgO | $Na_2O$ | 烧失量 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 154 | 徐深1井 | 3532.23 | K1yc | 流纹质角砾岩 | 0 | 0.066 | 0.286 | 0.463 | 4.042 | 0.013 | 0.063 | 69.971 | 13.031 | 0 | 4.921 | 2.226 |
| 155 | 徐深1井 | 3632.25 | K1yc | 火山角砾岩 | 0 | 0.071 | 0.221 | 0.109 | 4.678 | 0.013 | 0 | 78.876 | 13.276 | 0 | 4.698 | 1.426 |
| 156 | 徐深1井 | 3634.80 | K1yc | 火山角砾岩 | 0 | 0.069 | 0.214 | 0.101 | 3.925 | 0.014 | 0.003 | 76.344 | 12.731 | 0 | 4.850 | 1.486 |
| 157 | 滨参1 | 2000.50 | J3dr | 泥岩 | 4.294 | 0.033 | 0.437 | 0.723 | 4.055 | 0.021 | 0.069 | 63.221 | 17.253 | 0.897 | 1.700 | 7.208 |
| 158 | 滨参1 | 2006.00 | J3dr | 泥岩 | 5.577 | 0.054 | 0.469 | 0.750 | 3.606 | 0.034 | 0.059 | 62.434 | 16.440 | 1.008 | 1.856 | 7.631 |
| 159 | 滨参1 | 2141.50 | J3dr | 泥岩 | 4.823 | 0.040 | 0.547 | 0.959 | 3.992 | 0.026 | 0.068 | 63.778 | 16.548 | 1.456 | 1.919 | 5.720 |

**Figure 1.** Elemental data overview
**图 1.** 元素数据概览

从图 1 可见，每一个样品我们共分析了 11 个常量元素，这 11 个常量元素就基本占到样品里 90%以上的含量，是地质应用常用的分析数据。但是在数据里出现了值为 0 的情况，这可能是因为其元素含量在百万分之一以下从而导致荧光光谱仪无法正常的测量。鉴于之后我们要用到 K 均值聚类法进行数据分类，值为 0 的数据会造成分类偏差，所以数据处理的第一个步骤就是要引入“过滤”算子，过滤掉所有存在数值为 0 的样品。

另外从岩石样品的岩性上来看，数据里存在着 100 多种不同种类的岩石分类，在此我们选用泥岩作为分析的指标，故而数据处理的第二个步骤是在“过滤”算子只留下岩性为泥岩的样品数据，见图 2。

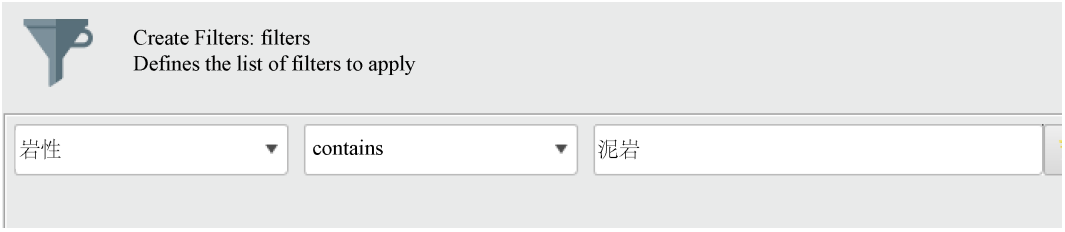


**Figure 2.** Filter operator
**图 2.** 过滤算子

经过以上两步的数据过滤后，我们提炼出 2719 块泥岩样品的数据，然后就可进行 K 均值聚类法进行数据分类。

## 3.2. K 均值聚类法算法原理

在数据处理完毕以后，我们希望对泥岩样品做个简单的分类概况，这里将用到 K 均值聚类法。

K 均值聚类属于基于原型的聚类方法，数据集会聚出 k 个簇，它也是是最简单最常用的聚类算法之一。该算法将指定用户事先设定簇的个数，目标是找出各簇的质心(centroid)，而后与各质心相邻的数据点聚成各簇，从而实现聚类。簇的质心可以是该簇内所有点的均值，即进行 k 均值聚类；也可以是中位点，即进行 k 中心点聚类。质心并不一定是真实存在的数据点，也可以是虚拟的点，只要它能最大程度的代表这个簇就行。其原理非常简单，对于给定的数据集总能有解。不过多数情况下这个解只是局部最优，并非全局最优。

K 均值聚类的计算过程是先初始化 k 个质心，然后找出距离最近质心的各个数据点，从而形成个簇。

确定每一簇以后，就可以计算各簇的新质心。然后更新各个数据点至最近的质心，确定新的簇并再一次更新质心，直到各数据点不再变更自己所属的簇或者这个变更不再显著为止。这样最后确定的质心就是数据内部各簇的代表或者原型，它们可以描述整个模型。

K 均值聚类的优点是简单、易于实现、结果易懂。虽然该算法的确可以处理高维数据集，但多次迭代和运行的计算成本还是很高的。最好的方法是一开始时把 k 值设的比较小，而后慢慢增大直到拟合[6]。

### 3.3. 流程设计

RapidMiner 提供了很方便的可视化流程设计界面，用户只需要简单的拖曳算子即可完成，本次为泥岩分类的整个流程图见图 3。

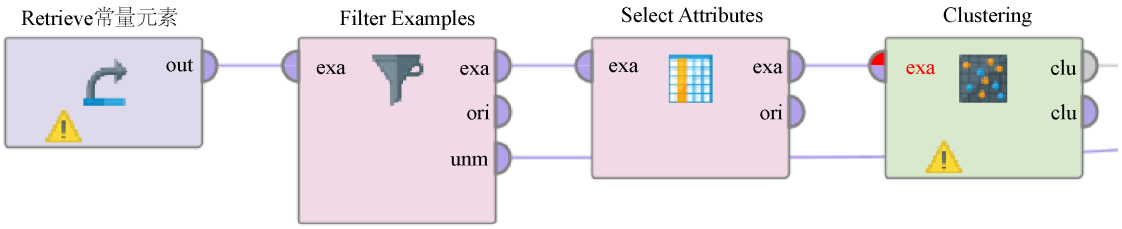


**Figure 3.** Data processing flow
**图 3.** 数据处理流程

其中各个算子的作用如下：

“Retrieve 常量元素”：即未经过处理的原始样品常量元素的数据集，可以是各种数据库格式，本例里为 excel 格式。

“Filter Examples”：过滤算子，我们通过设置必要的参数过滤掉数值为 0 以及岩性为非泥岩的样品。

“Select Attributes”：选择属性算子，我们只选择了所有数值型的泥岩元素数据方便 K 均值聚类算法进行计算分类。

“Clustering”：分类算子，内含 K 均值聚类算法。

### 3.4. 运行与解析

把评估算子的输出与总结果借口连接起来，就可以保存并运行该流程了。其运算过程仅花费了 3 秒的时间，很高效。运行的结果有 2 种表达的方式，第一种见图 4。

**Cluster Model**

Cluster 0: 142 items
Cluster 1: 2577 items
Total number of items: 2719

**Figure 4.** Classification results of mudstone
**图 4.** 泥岩分类结果

即为分类的概况总结，从 2719 块泥岩样品从通过 K 均值聚类法得到两种分类以及每种分类的数量。而 RapidMiner 亦会提供每一类的元素平均含量结果见图 5。

从以上两种图解可以分析出，cluster_0 类(计 142 块样品)为富钙的钙质泥岩；cluster_1 类(计 2577 块样品)为富硅的硅质泥岩。

另外仅选用“过滤”算子我们从颜色上进行分类，发现松辽盆地的泥岩也可分为两类，其中红、褐色泥岩 11 块，暗色(黑色、灰色等)泥岩 2708 块。

| Attribute | cluster_0 | cluster_1 |
|---|---|---|
| $Fe_2O_3$ | 5.502 | 5.260 |
| Mn | 0.169 | 0.092 |
| Ti | 0.578 | 0.493 |
| CaO | 11.134 | 1.221 |
| $K_2O$ | 2.368 | 3.385 |
| S | 0.443 | 0.210 |
| P | 0.170 | 0.114 |
| $SiO_2$ | 50.952 | 64.789 |
| $Al_2O_3$ | 15.523 | 13.760 |
| MgO | 2.508 | 1.183 |
| $Na_2O$ | 1.786 | 1.800 |

**Figure 5.** Average element content of two kinds of mudstones
**图 5.** 两类泥岩的元素平均含量

## 4. 结论

1) 松辽盆地的沉积填充中，泥岩约占 58%，且主要构成为硅质泥岩，其主量元素的平均含量为氧化铁 5.260%、锰 0.092%、钛 0.493%、氧化钙 1.221%、氧化钾 3.385%、硫 0.210%、磷 0.114%、二氧化硅 64.789%、三氧化二铝 13.760%、氧化镁 1.183%、氧化钠 1.800%。

2) 松辽盆地泥岩多为暗色系，说明盆地是在还原环境中沉积而形成的。

## 参考文献

[1] 王雅春, 朱焕来. 油气数学地质[M]. 北京: 石油工业出版社, 2015: 1.

[2] 谭永杰, 文敏, 朱月琴, 等. 地质数据的大数据特性研究[J]. 中国矿业, 2017, 26(9): 69.

[3] 康新荣. 含油气系统中泥岩多样性的成因及其对烃源岩、盖层、储集岩特性的影响[J]. 石油科技动态, 2012(6): 58.

[4] 乐昌硕. 岩石学[M]. 北京: 地质出版社, 2005: 102-105.

[5] 荀文婧, 徐铭明, 刘晓峰, 等. 森林大气温度、大气湿度与光照的关联研[J]. 电脑知识与技术, 2017, 13(22): 209.

[6] Vijay. 预测分析与数据挖掘 RapidMiner 实现[M]. 严云, 译. 北京: 人民邮电出版社, 2018: 182-190.