1. 四列聚类

因为数据中心很多条将地点与故障设备重复，该聚类的结果有按设备分类的倾向。大概分类效果如下：

19. 六轴数控叶片磨床-杂项
18. 小法拉利-刀
17. 小法拉利-油液
16. 力西提-杂项（以各种报警居多）
15. 力西提-漏气
14. 电气相关故障
13. 小法拉利-各种轴故障
12. 漏油漏液故障
11. 长征-刀与其他杂项
10. 数控铣床-杂项
9. 德玛吉与马扎克-杂项（刀、轴、管破裂）
8. 五轴叶片加工中心-杂项
7. 起重机故障
6. 小法拉利（数控五坐标加工中心）-杂项
5. 五坐标加工中心的各种机床-杂项
4. 五坐标加工中心-刀
3. 振动塞、砂管故障
2. 漏气故障
1. 法拉利（不是小法拉利）故障-杂项
0. 哈米乐（高速五坐标加工中心）故障-杂项

因为数据一部分是重点描述故障地点与机器，一部分是重点描述故障现象，所以这次聚类的结果也表现为一部分类以故障现象聚类在一起（如类18，类12），一部分类以故障设备聚类在一起（类19，类7，类6，类1），还有一部分则结合了上面二者（类18，类17，类4）.

2. 两列聚类与三列聚类

只对描述和故障现象两列进行聚类的话就减少了上面提到的故障地点的权重，但是由于数据中对于故障现象的描述很不精确，有些描述甚至对于聚类完全没有意义。观察来看，聚类效果并不是很好，很难使用关键词去概括。

三列聚类的效果和四列聚类基本一样，相比较于两列聚类的效果，加强了故障设备在聚类中的权重。

3. 评价

我认为如果在使用场景中用户可以大致判断是哪个设备出问题的话，四列聚类的效果会更好一些。对于四列聚类得到的20小类，其还可以主要分为铣床（法拉利、哈米乐、德玛吉、马扎克），磨床，起重机，电机与程序，油气管道这几大类，大类中的故障现象基本相似。可以考虑如对于铣床类，将上面得到的20小类中的法拉利、哈米乐、德玛吉、马扎克再合成为一个大铣床类，以期得到铣床类的故障共性。

还有就是数据中心有很多几乎没有价值的数据，后面可以考虑无论是在聚类、大模型还是给东电那边让工程师提供解决方案时都可以将这些垃圾数据去除掉。