设备维修信息数据挖掘

摘　　要

随着市场竞争的日益激烈，维修售后服务成为了企业的重要竞争能力之一。然而由于产品故障的不确定性使得备件需求难于预测，维修备件越来越多使得备件库存维护成本不断增加。这些问题使得维修企业面临的负担加重。因此针对产品的备件需求问题，本文利用某设备生产企业的维修数据记录，基于数据挖掘技术对不同型号的手机常见故障进行分析，从而为公司的设备储藏提供意见。

首先，本文对原始维修数据记录进行了简单分析。在对噪声数据和“服务商代码”进行预处理之后，将数据集中的手机维修信息提取出来。接着利用 clementine12.0 软件分析得知“反映问题描述”属性与手机使用时长、市场级别、服务商所在地区、产品型号相关性较强。

其次，为了分析故障与其他属性的关系，本文采用关联规则 Apriori 和 GRI 算法分析手机使用时长、产品型号分别与故障之间的关联性。观察关联结果，发现最近买的手机（使用时间低于两个月）主要故障集中在 LCD 显示故障和网络故障；较早买的手机主要出现开机故障和通话故障。但是 GRI 算法得出的结果支持度或置信度较低，不具有说服力。所以本文主要利用基于协同过滤的推荐算法来分析反映问题描述属性与其他属性的关联规则，并得出了如下结果：地理位置上相近的地区，其手机常见故障也类似；不同种手机型号或不同地区的手机出现的常见故障都是：开机故障，触屏故障，按键故障和通话故障；在不同级别的市场购买手机,，其经常出现故障的手机的手机型号都是 T818，T92，EG906，T912 和 U8。

最后，为了验证推荐算法的可信性，本文对该算法进行质量评价，利用 Celmentine 将数据分为训练集和测试集，然后进行算法检验。结果表明，推荐算法能够比较准确地得出推荐结果。

关键词：设备维修、clementine12.0 软件、GRI 算法、基于协同过滤的推荐算法

Data mining of equipment maintenance information

Abstract

As the competition in the market is increasing, maintenance after-sale service becomes one of the important competition ability of enterprise. However, due to the uncertaint breakdown of product, the spare parts demand is difficult to predict. And with the emergence of a growing number of maintenance spare parts ,the cost of Inventory maintenance is increasing. All of these problems make maintenance enterprises are faced with the burden. Therefore, aiming at Spare parts demand for the product, we use the maintenance record of a equipment manufacturing enterprise to analyse common breakdown of different kinds of mobile phones based on data mining technology and provide equipment storage advices to the mobile phone company.

First of all, the article analyses the original maintenance data records. After preprocessing the noise data and 'Service providers code', we extract the data set of mobile phone repair information. Then we use clementine12.0 software to analyse the correlation between the properties and learn that 'The description of reflecting problem' has a strong correlation with 'The usage time of mobile phone' , 'The market level', 'Service area' and 'Product model'.

Then, In order to analyze the correlation between 'The description of reflecting problem'and other attributes, We use Apriori and GRI algorithm to analyze the correlation between 'The description of reflecting problem' and 'The usage time of mobile phone' , 'Product model'. Observing the correlation results,we find that the breakdown or the cellphone bought within a month is focused on the LCD display and Network fault,and the cellphone buy early appears starting up fault and communication falut mainly.However, the support or confidence of the results are so low that the results are not convincing. So we mainly use recommendation algorithm which is based on the collaborative fitering to analyse the correlation between 'The description of reflecting problem'and other attributes.Finally,we get the following results:

1.The geographical position which is close its mobile phone common faults is similar;

2. Although the product model or service area is different,the cellphone appears the same following common faults: starting up fault , touch screen fault, button fault and communication falut;

3. Although the market level is different, the cellphone which appear fault usually is T818，T92，EG906，T912 和 U8.

Finally, in order to verify the credibility of the recommendation algorithm, this article is to evaluate the quality of the algorithm.The data is divided into training set and test set used Celmentine, and then test the algorithm. The results show that, the recommendation algorithm can obtain more accurate recommendation results.

Key: Equipment maintenance,Clementine12.0 software,The GRI algorithm,The recommendation algorithm which is based on the collaborative fitering

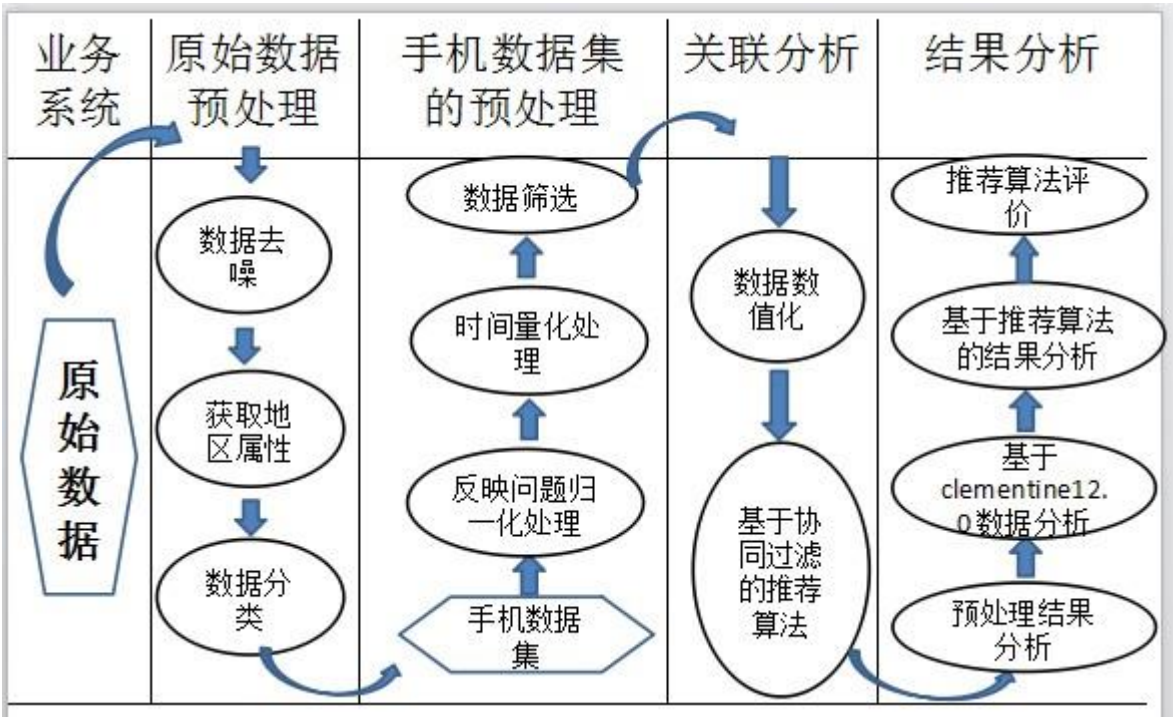
目录

1. 挖掘目标

本次建模目标是利用维修记录的海量真实数据，采用数据挖掘技术，分析手机各类故障与手机型号、手机各类故障与市场的相互关系，构建反映各类型号手机的常见故障评价指标体系、不同市场和地区手机质量的评价体系，为手机公司的设备储藏提供意见，同时也可为消费者提供购买意见。

2. 分析方法与过程

2.1. 总体流程



本文主要包括如下步骤：

步骤一：维修数据集的特点分析

分析原始维修数据集的特点，发现客户要求服务类型中安装记录与维修记录条数持平；数据缺失严重部分主要集中在购买商场、购机价格、机型属性、工程单号、工程总数、多次维修内机编号、故障原因代码、故障原因描述、维修措施这几个属性；数据噪声严重的部分集中在购机价格、产品型号。

步骤二：数据的预处理及筛选

首先对数据进行去噪处理，并且根据原始数据的“服务商代码”属性进行“服务商所在地”属性的提取。接着对数据集中的各项数据进行统计分析，为了方便进行更深入的研究，选取手机数据集进行进一步的数据挖掘研究。同样需要对手机数据集进行进一步的数据预处理，将“反映问题描述”属性的属性值进行归一化处理，根据购机日期属性及预约日期属性提取手机使用时长。并且利用 clementine12.0 软件分析得知“反映问题描述”属性与手机使用时长、市场级别、服务商所在地区、产品型号相关性较强。

步骤三：利用基于协同过滤的个性推荐算法进行关联分析

基于预处理后的高质量的手机数据集，采用基于协同过滤的个性推荐算法对手机数据集中的属性进行关联分析，并对该设备生产企业的备件储备需求进行预测分析、潜在故障预警分析、易损件及原因分析等方面进行探索分析。

2.2. 具体步骤

2.2.1. 维修数据集的特点分析

设备生产企业伴随着销量的增加，维修也在不断增加，随着时间的推移，越来越多的维修记录被存储到数据库中。当这些数据量积累到一定程度时，必然反映出一定的规则。但是在记录维修数据时，由于人工操作的失误以及客户的遗忘，使得了维修数据集存在缺失、噪声，而这些数据又影响着最终的结果。为了得出比较精确的决策，需先对数据集进行分析处理。从数据库中导出某设备生产企业的 685413 条维修记录数据，从该维修数据集分析可以得出以下特点：

(1)　每条维修记录提供了 29 个属性，属性的基本说明如表 1 所示：

表 1　　属性的基本说明

属性	属性描述
购机日期	客户购买商品的时间 （ 年/月/日）
购买商场	客户购买商品的商场
购买价格	客户购买商品的价格
机型属性	例如：节能惠民
市场级别	一类地区、二类地区、三类地区、四类地区
安装日期	设备安装日期（年/月/日）
预约日期	客户预约时间 (年/月/日)
信息编号	维修记录信息编号

工程单号	维修工程序列
工程单	是/否
工程总数	同一工程单号的工程数
多次维修	是/否
产品大类	TK 特种空调、冰箱、电视、机顶盒、家用空调、冷柜、手机、特种空调、洗衣机、专业冷柜
品牌	澳柯玛、哈士奇、华宝、康拜恩、西门子
产品型号	例如：LED42EC260JD
序列号	例如：AK0072014W0010NCQPD0282
内机编号	维修设备的序列号，例如： 1DBC29SW0CNG03M4R360765
服务商代码	例如：H-TV-532-0027 TV 代表电视机 532 代表区号
受理时间	（年/月/日/时/分/秒）
派工时间	（年/月/日/时/分/秒）
故障原因代码	例如：HJDYY91000 JD 代表机顶盒
故障原因描述	例如：交流接触器线圈短路
维修措施	例如：更换交流接触器
反映问题描述	例如：不制冷
要求服务类型	安装、换机、鉴定、调试、退机、维修、咨询
要求服务方式	例如：登门维修
实际服务类型	TF、安装、换机、鉴定、调试、退机、维修
实际服务方式	例如：登门维修
保修类型	例如：保外 保外代表保修期外

(2) 对数据集中的要求服务类型进行统计，客户要求服务类型为维修的数据记录有 314132 条，而安装的数据记录有 302824 条，与维修记录条数持平。因此可以得出该维修数据集不仅仅是商品需要维修的维修记录，而是商品售后服务的记录。

(3) 对数据集中的预约时间进行分析，可以发现预约时间全部在 2013-9 内，因此我们所获得的数据集是该设备生产企业 9 月份中客户预约维修的维修记录，即设备生产企业一个月内的维修工作。

(4) 数据集中各项属性存在部分或大量的数据缺失，缺失严重部分主要集中在购买商场、购机价格、机型属性、工程单号、工程总数、多次维修内机编号、故障原因代码、故障原因描述、维修措施这几个属性。

(5) 数据集中存在噪声数据，噪声严重的部分集中在购机价格、产品型号。由于客户忘记了当初的购机价格、填写价格错误等原因，该数据集中有 484164 条数据记录显示购机价格为 0。同样，可以认为由于工作人员登记方式不同，使得产品型号中存在大量噪声数据---型号的末端存在异常，例如“BCD-398WT-J，”。

2.2.2. 维修数据集的预处理

（1）缺失数据的处理

由于数据集中的数据量大，同时数据中的缺失部分规律不明显，无法找到合理的补充数据规律，那么在数据缺失严重的属性中随意或有限度的补充数据不仅会影响数据的筛选甚至会影响最终的结论。因此在该步骤中不对缺失数据进行处理，而是在后面的处理中针对不同问题对缺失数据进行处理。

（2）噪声数据的处理

利用 excel 表格对数据集中产品型号的错误数据进行修正，去除末端存在异常的型号的末端：选择产品型号——数据——分列——分隔符号——逗号。

由于型号与价格的关联不明显，而且有 484164 条价格记录是由于客户忘记当初的购机价格而填写为 0，从剩余的价格记录中难于寻找合理的规律修正错误的价格，因此不对错误价格进行修正。

（3）服务商所在地区的提取

数据集中并没有直接提供服务商的所在地，即提供产品售后服务机构的所在地，但是可以从服务商代码中提取出服务商所在地。利用 excel 表格提取服务商所在地，由于区号在服务商代码的第二个‘-’和第三个’-’之间，因此对于服务商所在地的提取公式为：

MID(服务商代码,FIND("-",服务商代码,3)+1,3)

例如：excel 表格中 R58 的服务商代码“H-SJ-010-0014”，提取公式 MID(R58,FIND("-",R58,3)+1,3)可得到 010。

010 代表服务商所在地的区号，通过网上查找可得 010 是北京的区号，即可得出该服务商所在地为北京。

（4）数据的筛选

从数据的分析中可知道数据集中含有的 7 种服务要求类型以及 10 种产品大类，从数据集中筛选出安装数据记录，其余的可归为维修记录。利用 excel 表格做分析如下：

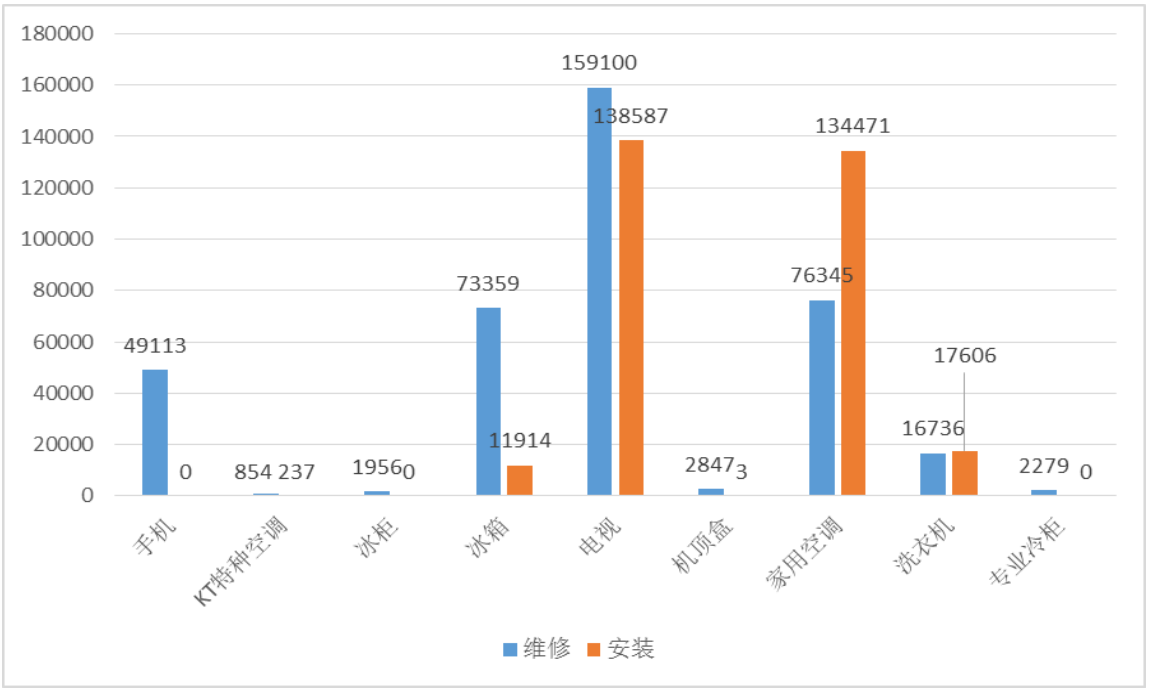


图 1 维修和安装记录分布

由于不同产品其隐含的规则不同，不能全部一起讨论，需先逐一分析各个产品，而分析方法都相同，不需要对全部进行讨论，只需提取其中的一种产品进行详细的分析，其他产品的分析过程类似。通过观察图 1，得知手机只有维修记录，并且记录数不少，便于分析挖掘其中隐含的规则且不失一般性，因此下面将选择手机进行深入挖掘研究。

（5）手机数据集的预处理

在数据挖掘整体过程中,海量的原始数据中存在着大量杂乱的、重复的、不完整的数据，严重影响到数据挖掘算法的执行效率，甚至可能导致挖掘结果的偏差。为此，在数据挖掘算法执行之前，必须对手机数据集进行预处理，以改进数据的质量，提高数据挖掘过程的效率、精度和性能。

1） 反映问题描述属性的归一化处理

由于数据集中反映问题的描述没有一定的标准，因此对于相同的手机故障，客户反映问题的描述也是各样，例如针对手机白屏问题，就有“开机白屏”和“白屏”这两种不一样的表示方式。不一样的标准不仅影响其他属性与反映问题描述属性进行关联规则分析，更严重影响了研究手机某型号设备出现的常见故障现象，因此需要对反映问题描述属性进行归一化处理。

根据手机故障原因标准准则（见附录 1），利用 excel 表格通过筛选包含相同字样的数据记录，并进行替换。例如：

反映问题描述	故障描述
开机白屏	LCD 显示故障
白屏	LCD 显示故障
开不了机	开机故障

2）客户手机使用时长的量化处理

根据多次维修属性的反映，每条记录均为否或空白，因此可以假设客户的手机均为第一次维修。观察数据发现购机日期有 430 个缺失数据，而最早的购机日期为 2011-10-10，故使用 2011-10-01 填补空缺数据。将预约日期减去购机日期所得即为客户手机使用时长。

3）手机数据集属性的筛选

由于本文的研究目的在于研究手机常见故障与手机型号、手机各类故障与市场的相互关系，而原始数据共有 29 个属性，所以为了提高算法的的效率和精度，本文只需要对购机日期、预约日期、市场级别、服务商代码、产品型号、反映问题描述进行分析即可。由于服务商代码、产品型号、反映问题描述存在缺失数据，而缺失部分规律不明显，无法找到合理的补充数据规律，所以在此采用忽略缺失数据的处理方法。

将归一化处理后的“反映问题描述”属性与手机使用时长(上述购机日期与预约日期的关联后得到的数据)、市场级别、服务商所在地区(从服务商代码中提取出来的)、产品型号 4 个属性进行变量重要性分析。使用 clementine12.0 软件，通过 Modeling 卡中的 Feature Selection 节点对上面的属性进行变量重要性的分析后得知“反映问题描述（以下简称故障）”属性与手机使用时长、市场级别、服务商所在地区（以下简称地区）、产品型号相关性较强。（见图 2）

图 2 变量重要性分析

2.2.3. 关联分析

随着数据库中的数据量急剧增长，人们获取自己感兴趣的信息越来越困难，面对海量的数据，每个用户希望获取的信息可能仅仅是其中很少的一部分，同时用户的需求常常是模糊的、不明确的，可能会对某些信息存在着潜在的喜好。如果服务提供者能够把信息推荐给用户，就可能把用户的潜在需求变为现实进而盈利。在这种背景下，推荐系统应运而生，推荐系统的实现方法众多，其中基于协同过滤推荐算法理论上可以推荐世界上的任何一种东西，适用性强，也是一种最成功的推荐系统算法。

文本上述研究表明手机使用时长、市场级别、地区、产品型号与故障属性都具有关联，但是各属性的对象很多，而且手机维修记录数目也达到了 47000 多条，这导致了设备生产企业难于分析备件的常见故障及潜在故障预测。查找文献了解到 Apriori 算法基于关联规则的推荐算法的基本算法，利用 clementine12.0 软件进行基于 Apriori 算法的关联规则分析，但是发现结果的支持度及置信度较低。

基于协同过滤的个性推荐算法可以很好的解决该类问题。通过推荐算法可以分析反映问题描述属性与其他属性的关联规则，产生推荐项目，从而构建反映各类型号手机的常见故障评价指标体系、不同市场和地区手机质量的评价体系，为手机公司的设备储藏提供意见，同时也可为消费者提供购买意见。因此本文选择基于协同过滤的个性推荐算法建立模型，进行属性间的关联分析挖掘，完成推荐。

（1） 数据数值化

通过上述的数据处理步骤后得到的手机数据集，包括以下五个属性：使用时长（数值型）、市场级别(字符型)、产品型号（字符型）、地区（字符型）、故障（字符型）。为了方便进行基于协同过滤的个性推荐算法的数据挖掘，需要将属性的数据均数值化。可利用 excel 表格手动将上述属性进行数值化处理，如市场级别属性的处理，其他属性均可如此处理，但是产品型号的类别较多，手动处理耗时，可利用 Matlab 编写函数文件对其数值化。（程序代码见附录 2，属性所对应的数值化可见附录 3 及附录 4）

表 2 市场级别数值化对应值表

市场级别	一级市场	二级市场	三级市场	四级市场
数值化	1	2	3	4

表 3 故障数值化对应值表

反映问题描述	GPRS	LCD显示故障	MP3、收音故障	不读卡	充电故障	其他	喇叭故障	外观故障	开机故障
数值化	1	2	3	4	5	6	7	8	9
反映问题描述	拍照故障	按键故障	振动故障	灯故障	电池故障	网络故障	蓝牙故障	触屏故障	通话故障
数值化	10	11	12	13	14	15	16	17	18

（2）基于协同过滤的个性推荐算法的描述

1） 伪用户_项目评分矩阵的构建————用户偏好描述

收集有关用户偏好信息，如用户对商品的评分，通过对原始数据进行清理、转换和录入，最终形成一个 m×n 维矩阵。其中行代表用户，列代表项目。如下表 2 所示，$R_{i,u}$ 表示第 i 个用户对项目 u 的评分值,评分值为数值型。

	项目 1	……	项目 u	……	项目 n
用户 1	$R_{1,1}$	……	$R_{1,u}$	……	$R_{1,n}$
……	……	……	……	……	……
用户 i	$R_{i,1}$	……	$R_{i,u}$	……	$R_{i,n}$
……	……				
用户 m	$R_{m,1}$	……	$R_{m,u}$	……	$R_{m,n}$

表 4

2） 相邻用户矩阵的构建-----寻找最近邻居

在评分矩阵中，用户已评分的项目为实际的评分值，未评分项目用 0 表示。如果用户评分被看作是 n 维项目空间上的向量，余弦相似性就是将用户的相似性通过向量间的余弦夹角度量。设用户 i 和用户 j 在 n 维项目空间上的评分分别用$\overrightarrow{R_i}$，$\overrightarrow{R_J}$表示,在用户 i 和用户 j 之间的相似性计算公式如下所示：

$$sim(i,j)=cos(i,j)=\frac{\overrightarrow{R_i}\cdot\overrightarrow{R_J}}{\|\overrightarrow{R_i}\|\times\|\overrightarrow{R_J}\|} \quad ①$$

3） 产生推荐

最近邻居集产生后，可计算目标用户对项目的预测评分值进行 Top-N 推荐。通过预测评分值搜索最近邻居而产生推荐，预测评分计算公式如下：

$$P_{i,y}=\overline{R_i}+\frac{\sum_{j\epsilon NN,y\epsilon N} sim(i,j)(R_{j,y}-\overline{R_j})}{\sum_{j\epsilon NN,y\epsilon N} sim(i,j)} \qquad ②$$

其中，$P_{i,y}$代表目标用户 i 对项目 y 的预测评分值；$\overline{R_i}$ 为用户 i 的平均评分值；$R_{j,y}$ 表示目标用户 i 的最近邻居集的用户 j 对项目 y 的评分。在此，目标用户 i 的最近邻居集用 NN 表示。按评分预测值$P_{i,y}$的高低排序产生推荐集。

（3） 建立各地区常见故障的预测推荐

协同过滤算法需要整理用户的评分数据、计算相似性、寻找最近邻居从而完成推荐。但是对于本文以手机数据集为记录集，研究地区属性与反映问题描述属性之间的关联获得各地区中最常见故障集而言，各地区对反映问题描述并没有显式的评分，这需要 Web 挖掘算法来获得隐式的评分数据。基于 Matlab，构建算法的步骤如下：

1）地区_故障评分矩阵的构建

地区与故障的关联可通过手机数据集中的记录条数反映。由于原始的手机数据记录集已表明该客户并没有多次手机维修记录，因此地区 i 对故障 y 的评分可用在地区 i 的故障集中故障 y 的记录条数表示。但是考虑到手机使用时长并不完全相同，其分布如图 6 所示。客户当月购买的手机当月便进行维修，可见该手机的发生故障的概率是很大的，因此对该条记录，维修次数可修正为手机数据集中最大的使用时长减去当前记录的使用时长。假设$R_{i,u}$代表地区 i 对故障 u 的评分值，T_w代表使用时长，那么$R_{i,u}=\underset{w\epsilon N}{Max}(T_w)-T_{d(d\epsilon i,u)}$。根据该方法计算各地区对所有故障的评分值，形成地区的评分矩阵R_{mn}。

2）寻找最近邻居并进行推荐

寻找地区的最近邻的关键是计算各地区之间的相似性。可利用余弦相似性的计算方法，即公式①，通过地区与故障之间的评分矩阵R_{mn}计算相似度得到各地区之间的相似值所组成的矩阵 sim(m，m)，并通过 acos()函数得到数值均处于 0 和 1 之间的矩阵。对于地区 i 而言，把计算出的所有相似值按照从小到大选出若干个相似值小于α的作为其最近邻居集。

得到各地区的最近邻居集后，利用公式②，可得到各地区对于所有故障的预测评分

矩阵P_{mn}，按照评分预测值$P_{i,y}$的高低对地区 i 产生故障推荐集，该集代表地区 i 常出现的故障。(程序代码见附录 7)

该模型具有较大的灵活性。不仅可以进行各地区的潜在故障预警分析，从而实现企业在各地区的备件储备需求预测，还可以按支持度大小反映各手机型号设备出现的常见故障，只需要将手机型号作为用户、故障作为项目，直接运行程序即可。同时，由于手机维修属于送修类型，客户购机的地区并一定就是服务商所在地 ，从而手机数据集中的市场级别属性与服务商所在地属性的关联不显著。因此，可利用该模型对市场级别属性与产品型号属性进行关联分析，预测各市场级别最常出现故障的手机型号。

2.3. 结果分析

2.3.1 预处理的结果分析

根据服务商代码属性提取出地区属性，并根据各地区归纳到各省份(见附录 3 表 2)。

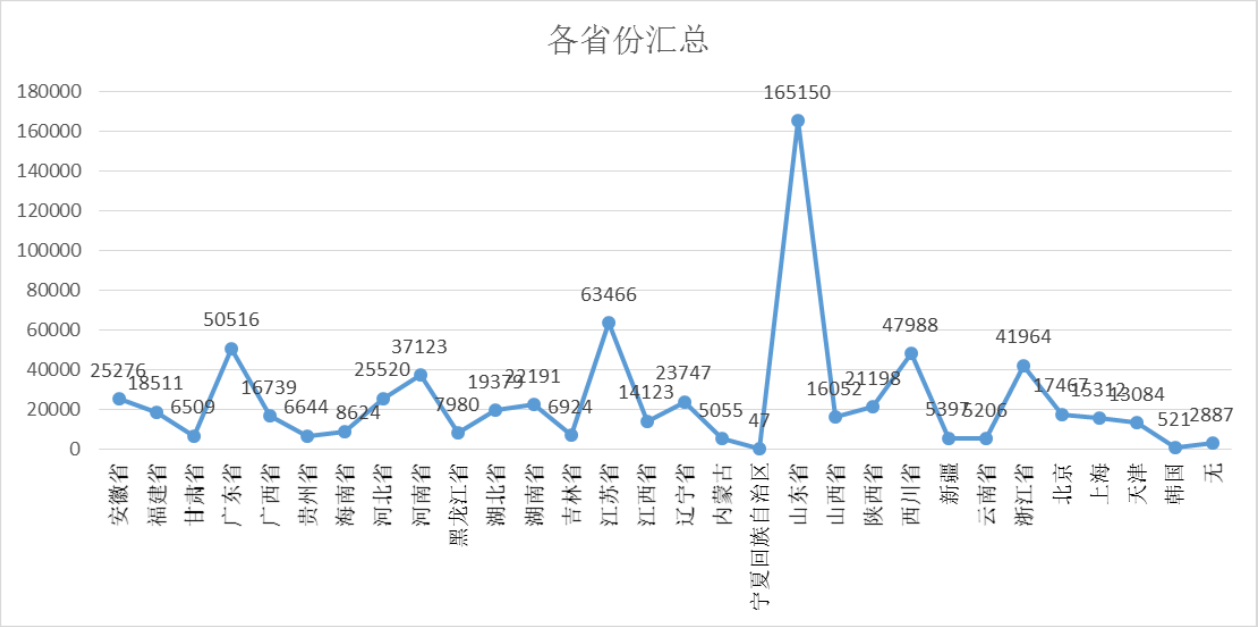


图 3 各地区的手机记录分布图

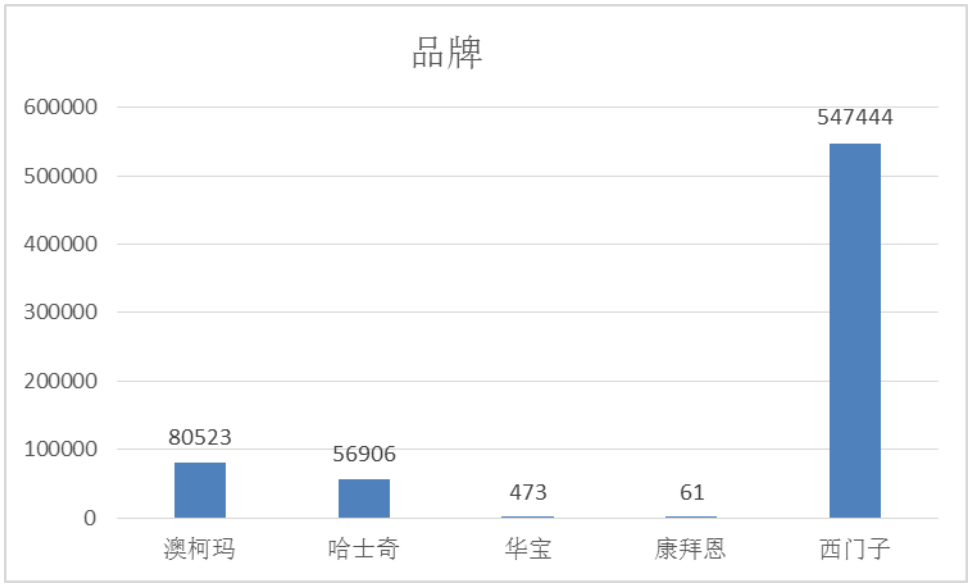


图 4 品牌分布情况

结论：从图 3 中可以发现维修记录基本囊括了全国各省，除西藏、清海省，表明了该设备生产企业是一个全国性的企业。同时山东省的维修记录数目远远大于其他各省份，可以认为山东省是该设备生产企业的总部所在地.从图 4 可以发现西门子的比重远远大于其他品牌，同样可以认为西门子是该企业的自身品牌。通过网上搜索查找，可以发现西门子本非山东省的本土品牌。因此可以认为该数据集中的品牌是经过某种规则映射而成的。

2.3.2 手机数据集基于 Clementine 结果分析

(1)基于 Apriori 的型号与故障关联分析

后项	前项	实例	支持度 %	置信度 %
故障描述 = 开机故障	产品型号 = T818	3,542	7.27	25.071
故障描述 = 触屏故障	产品型号 = T818	3,542	7.27	20.751
故障描述 = LCD显示故障	产品型号 = T818	3,542	7.27	17.42
故障描述 = 通话故障	产品型号 = T818	3,542	7.27	12.479
故障描述 = 按键故障	产品型号 = T818	3,542	7.27	7.595
故障描述 = 开机故障	产品型号 = T92	3,240	6.65	33.889
故障描述 = 触屏故障	产品型号 = T92	3,240	6.65	14.444
故障描述 = 通话故障	产品型号 = T92	3,240	6.65	11.883
故障描述 = LCD显示故障	产品型号 = T92	3,240	6.65	10.37
故障描述 = 喇叭故障	产品型号 = T92	3,240	6.65	6.451
故障描述 = 振动故障	产品型号 = T92	3,240	6.65	5.586

图 5 基于 Apriori 的型号与故障关联

按手机型号的支持大小，可以发现型号 T818 的支持度最高，但是仅仅达到 7.27%，而其

最常见的故障分别为开机故障、触屏故障、LCD 显示故障、通话故障、按键故障。对于支持度为 6.65%的 T92 而言，最常见故障如图 5 所示。

(2)使用时长属性分析

1）使用时长，产品型号与故障的散点图

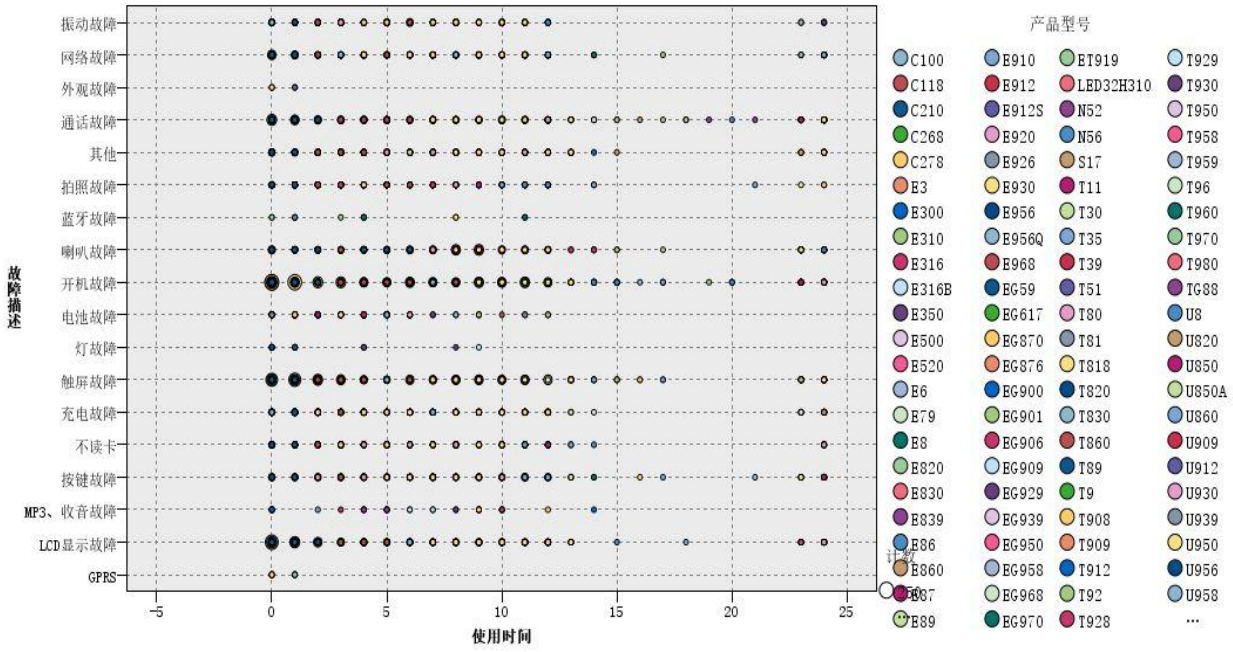


图 6

结论分析：图 6 中散点的大小表示数量的多少，可看出不同种手机型号的手机故障都主要集中在开机故障，LCD 显示故障，触屏故障和通话故障；在短时间内进行手机维修的比较多，而使用时间较长的维修数比较少。因为现在手机更新快，价格合理，同时跟据人类消费心理，当顾客购买的手机在短时间内出现故障问题，基本会选择维修，但是如果手机使用时间长并且出现了故障，顾客就会选择直接换手机，而不是维修。因此在手机的维修数据集中手机已使用时长较短时长的比重大，使用时间长的数据偏少。

2）手机使用时长与故障的 GRI 关联分析

后项	前项	实例	支持度 %	置信度 %
故障描述 = 喇叭故障	使用时间 < 3.500	24,140	49.54	5.19
故障描述 = LCD显示故障	使用时间 < 1.500	16,629	34.13	16.71
故障描述 = 网络故障	使用时间 < 1.500	16,629	34.13	5.98
故障描述 = 按键故障	使用时间 > 6.500	15,307	31.42	6.09
故障描述 = 触屏故障	使用时间 > 7.500	12,450	25.55	22.35
故障描述 = 开机故障	使用时间 > 20.500	367	0.75	16.89
故障描述 = 充电故障	使用时间 > 22.000	364	0.75	5.22
故障描述 = 通话故障	使用时间 > 23.500	285	0.58	21.05

图 7

结论分析：虽然支持度或置信度不是很高，但也能够大概反映一些信息：最近买的手机(使用时长在两个月内)主要的故障集中在 LCD 显示故障和网络故障；使用时间比较长的手机主要故障集中在通话故障，充电故障和开机故障。

2.3.3 基于推荐算法的手机数据集分析

(1) 各地区常见故障的预测结果分析

1）地区_故障的评分值矩阵R_{mn}

根据R_{mn}，利用 Matlab 软件，运行程序（见附录 6）生成地区与故障的评分图如下所示。其中 x 轴代表数值化的故障，y 轴代表地区 i 对故障 u 的评分，不同颜色的线段代表各地区。从图中可以看出各地区在 x=2,9,15,17，即 LCD 显示故障、开机故障、网络故障、触屏故障，比较集中，并且各地区对所有故障的数量趋势一样。其中山东省的地区 531 的开机故障尤其突出。

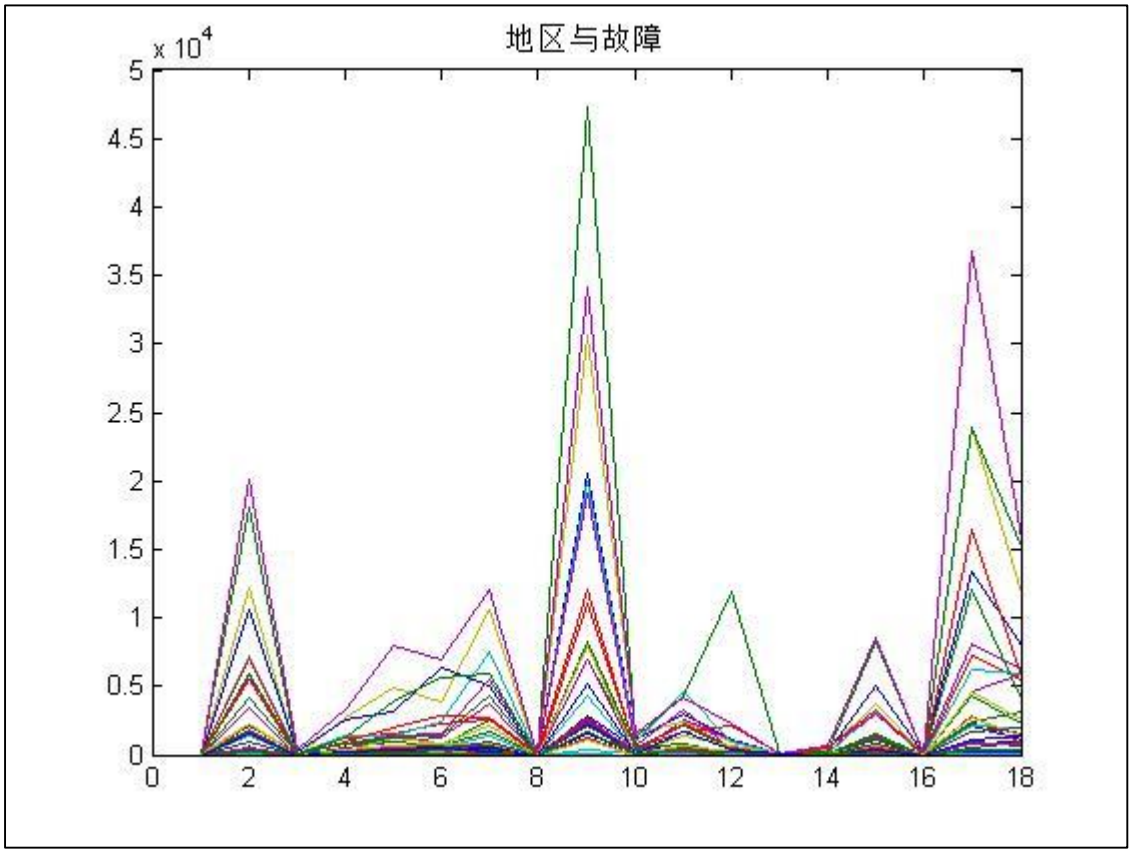


图 8

2）地区的最近邻 sim(m,m)

当取α=0.3，各地区产生的最近邻如下表格所示，当取α=0.2 时，各地区产生的最近邻如下表格的红色字体部分所示。由表格可以发现，当α=0.2 时，通过查看地图（见

附录 5）可以发现各地区的最近邻从地理位置上而言也是其近邻，这表明了地理位置上相近的地区具有相似性。因此地理位置上相近的地区，其常见手机常见故障也类似。

表 5 地区近邻表

地区	邻近值												
10	29	27	871	371	351	791							
21	25	28	791	311	531	371	731	571	871	27			
22	851												
23	29												
24	898	311											
25	21	571	28	791	311	731	371	755	551	531			
27	371	29	351	871	931	10	791	531	991	21	28		
28	311	731	25	21	791	571	591	531	371	27	771		
29	351	27	371	10	791	23	851	931					
311	28	531	731	791	25	21	571	591	371	898	24		
351	371	29	851	27	791	451	871	10					
371	791	351	27	851	29	871	21	531	25	451	28	311	10
431													
451	791	371	351	851									
471	991	931	871										
531	311	21	871	371	28	791	27	25					
551	571	755	771	731	25								
555													
571	551	731	755	25	771	591	28	311	21				
577	755												
591	731	571	898	28	5111	311	771						
731	591	571	28	771	311	551	25	21	755	5111			
755	571	551	25	771	577	731							
771	551	731	571	755	591	28							
791	371	451	351	311	851	25	28	21	29	27	531	871	10

851	371	351	791	22	451	29				
871	27	371	991	531	931	471	10	351	21	791
898	24	591	311							
931	991	27	471	871	29					
991	931	871	471	27						
5111	591	731								

3）各地区常见故障预测

运行程序（附录 7）观察结果，可得到如下结论：每个地区的手机故障主要是：开机故障，触屏故障，LCD 显示故障和通话故障。同时这四类手机故障在手机数据集中所占的比重是很大的，而根据推荐结果分析，可以发现各地区的常见手机故障都是这几类故障，表明了该品牌的手机出现故障与各地区的关联较低。

表 6 各地区常见故障表

地区	常见故障																
10	9	17	2	18	7	15	6										
21	9	17	2	18	7	15	6	11	5	12	4						
22	9	17	2	18	7												
23	9	17	2	18	7												
24	9	17	2	18	7	15	6	11	5	12	4	10	14				
25	9	17	2	18	7	15	6	11	5	12	4	10	14	3	13	1	16	8
27	9	17	2	18	7	15											
28	9	17	2	18	7	15	11	6	5	12	4	10	14	3	13	1	16	8
29	9	17	2	18	7	15	6	11	5	12	4	10						
311	9	17	2	18	7	15	6	11	5	12	4	10	14	3	13	16	1	8
351	9	17	2	18	7	15	6	11	5	12	4	10	14	3	13	16	1	8
371	9	17	2	18	7	15	6	11	5	12	4	10	14	3	13	1	16	8
431	9	17	2	18	7	15											
451	9	17	2	18	7	15											
471	9	17	2	18	7	15											

531	9	17	2	18	7	15	6	11	5	4	12	10	14	3	13	16	1	8
551	9	17	2	18	7	15	6	11	5	12	4	10	14	3	13	16	1	8
555	9	17	2	18	7													
571	9	17	2	18	7	15	6	11	5	12	4	10	14	3	13	1	16	8
577	9	17	2	18	7	15	6	5	11									
591	9	17	2	18	7	15	6	11										
731	9	17	2	18	7	15	6	11										
755	9	17	2	18	7	15	6	11	5	12	4	10	14	3	13	16	1	8
771	9	17	2	18	7	15	6	11										
791	9	17	2	18	7	15	6	11	5									
851	9	17	2	18	7	15												
871	9	17	2	18	7	15	6	11	5	12								
898	9	17	2	18	7	15	6	11										
931	9	17	2	18	7	15	6	11	5									
991	9	17	2	18	7	15												
5111	9	17	2	18	7	15	6	11										

4)不同手机型号的常见故障预测

结论分析:按维修记录支持度大小对型号进行排序,其中支持度最大的手机型号为 T818,其次为 T92。手机各型号常见故障可如下表格所示。其中常见故障为开机故障、触屏故障、LCD 显示故障、通话故障，与上述的基于地区的手机常见故障的推荐所产生的结果相近，也表明了该品牌手机出现的故障与手机型号的差异关联性较小。同时观察下表，可以发现该结果与基于 Apriori 的型号与故障关联分析的结果存在差异，但是差异并不大。产生差异的原因主要是基于协同过滤的推荐算法考虑了各手机型号之间的相似性，产生了新的手机型号与故障之间关联。

表 7　　各型号常见故障表

型号	常见故障																	
T818	9	17	2	18	7	15	6	11	5	12	4	1	14	3	13	16	1	8
T92	9	17	2	18	7	15	11	6	5	12	4	1	14	3	13	16	1	8
EG906	9	17	2	18	7	15	6	5	11	4	12	1	14	3	13	16	1	8

U8	9	17	18	2	7	15	6	11	5	12	4	1	14	3	13	16	1	8
T912	9	17	2	18	7	15	6	5	11	12	4	1	14	3	13	1	16	8
T930	9	17	2	18	7	15	11	6	5	12	4	1	14	3	16	1	13	8
E956Q	9	17	2	18	7	15	11	6	5	12	4	1	14	3	13	16	1	8
E820	9	17	2	18	7	15	6	5	11	12	4	1	14	3	1	13	16	8
E926	9	17	2	18	7	15	11	6	5	12	4	1	14	3	16	13	1	8
EG950	9	17	2	18	7	15	6	11	5	12	4	1	14	3	13	16	1	8
E912S	9	17	2	18	7	15	6	11	5	12	4	1	14	3	13	1	16	8
E830	9	17	2	18	7	15	11	6	5	12	4	1	14	3	13	1	16	8
T830	9	17	2	18	7	15	11	6	5	12	4	1	14	3	13	16	1	8
U930	9	17	2	18	7	15	11	6	5	12	4	1	14	3	13	16	1	8
EG901	9	17	2	18	7	15	11	6	5	12	4	1	14	3	13	16	1	8
E930	9	17	2	18	7	15	6	11	5	12	4	1	14	3	13	16	1	8
E912	9	17	2	18	7	15	6	11	5	12	4	1	14	3	13	16	1	8
E920	9	17	2	18	7	15	11	6	5	12	4	1	14	3	13	16	1	8
U860	17	9	2	18	7	11	6	5	15	4	12	1	14	3	13	16	1	8
E860	9	17	2	18	7	15	11	6	5	12	4	1	14	3	13	16	1	8
T96	9	17	2	18	7	15	11	6	5	12	4	1	14	3	13	16	1	8
EG970	9	17	2	18	7	15	6	11	5	12	4	1	14	3	16	13	1	8
T950	17	9	2	18	7	15	11	6	5	12	4	1	14	3	13	1	16	8
EG909	9	17	2	18	7	15	6	11	5	12	4	1	14	3	13	16	1	8
EG870	17	9	2	18	7	6	11	5	15	12	4	1	14	3	13	16	1	8
T860	9	17	2	18	7	15	6	11	5	12	4	1	14	3	13	16	1	8
T820	9	17	2	18	15	7	6	5	11	12	4	1	14					
TG88	9	17	18	2	7	15	11	6	12	5	4	1						
E956	9	17	2	18	7	15	6	11	5	12	4	1						
EG929	9	17	2	18	7	15	6	11	5	12	4							
U950	9	17	2	18	7	15	6	11	5	12	4							
T929	9	17	2	18	7	15	6	5	11	12	4							
T958	9	17	2	18	7	15	11	6	5	12	4							
U850A	9	17	2	18	7	15	11	6	5									
T909	17	9	2	18	7	11	15	6	5									
EG958	9	17	2	18	7	15	11	6	5									
EG900	9	17	2	18	7	15	11	6	5	12								
E910	9	17	2	18	7	15	11	6	5	12								
T960	9	17	2	18	7	15	6	11	5									
U820	9	17	2	18	7	15	11	6	5									
EG876	9	17	2	18	7	15	6	11	5									
U850	9	17	2	18	7	15	6	11										

T908	9	18	2	17	7	15	11	6										
U912	9	17	2	18	7	15	6	11										
T970	9	17	2	18	7	15	6	11										
U960Q	9	17	2	18	7	15	6											
U970	9	17	2	18	7	15	6											
E89	17	9	2	18	7													
U958	9	17	2	18	7	15	11	6										
T30	9	7	18	2	17	15	6	5										
U909	9	17	2	18	7	15												
U939	9	17	2	18	7	15												
T89	9	17	2	18	7	15												
ET919	9	17	2	18	7	15												
T80	9	17	2	18	7													
E316	9	2	17	18	7	15												
EG939	9	17	2	18	15	7												
T39	9	2	18	5	7	15	17	6										
E310	9	18	17	2	7	15												
E520	9	2	18	17	7	15												
E968	9	17	2	18	7	15												
T35	9	2	18	17	15	7												
E87	17	9	2	18	7													
T81	9	17	2	18	7													
EG968	17	9	2	18	7													
E316B	9	7	18	17	2													
T959	9	2	17	18														
T51	1	2	3	4	5	6	7	8	9	1	11	12	13	14	15	16	17	18
E839	9	17	2	18	7													
C210	9	18	2	7	5	11												
E350	9	18	2	17	7													
T11	1	2	3	4	5	6	7	8	9	1	11	12	13	14	15	16	17	18
N52	9	7	17	18	2													
E300	1	2	3	4	5	6	7	8	9	1	11	12	13	14	15	16	17	18
T9	6	5	9	15	17	1	2	3	4	7	8	1	11	12	13	14	16	18
C278	9	18	2	17	7													
E86	9	18	17	2	7													
E6	6	15	2	4	9	17												
C100	9	18	17	2	7	15	11											
N56	9	11	18	7	2													
E500	9	18	2	17	7													

E8	9	18	2	17	7													
E3	17	9	2	18														
E79	9	18	2	17	7													
EG617	9	2	17	18	15													
T928	1	2	3	4	5	6	7	8	9	1	11	12	13	14	15	16	17	18
U956	9	1	17	2	18													
C118	9	18	2	17	7													
C268	1	2	3	4	5	6	7	8	9	1	11	12	13	14	15	16	17	18
EG59	9	18	2	17	7													
LED32H310	17	9	2															
S17	9	18	2	17	7													
T980	1	2	3	4	5	6	7	8	9	1	11	12	13	14	15	16	17	18

5）不同市场的常出故障的手机型号预测

结果分析：如下表所示，在不同级别的市场购买手机,，其出现故障较多的手机型号都是T818，T92，EG906，T912 和 U8 。综合上述地区与故障的关联分析及手机型号与故障的关联分析，为该设备生产企业的各地区服务点提供了备件需求预测。即该品牌手机常出现故障的手机型号为 T818，T92，EG906，T912 和 U8，而常见故障类型为开机故障、触屏故障、LCD 显示故障、通话故障，。同时图 3 已表明手机维修数据集中各省份各地区的维修记录数是存在差异的。设备生产企业需根据各地区的支持度大小及时向各地补给这些种类的备件。

表 8　市场级别与手机型号推荐关联性

市场级别	手机型号							
1	59	68	39	80	67	17	71	31
2	68	59	67	80	39	71	17	31
3	59	68	67	39	80	31	71	26
4	59	68	39	80	18	67	31	17
注：T818（59），T92（68），EG906（39），T912（67），U8（80）								

2.3.4 推荐算法的评价

（1）算法推荐质量评价方法的描述

推荐质量的评价标准有多种，做常用的是通过平均绝对偏差（MAE）对推荐质量进行评价。MAE 通过计算预测用户评分与实际的用户评分之间的偏差度量预测的准确性，MAE 越小，推荐质量越高。假设预测的用户的评分值合为$\{p_1,p_2,\cdots,p_n\}$,对应的实际评分集合为$\{q_1,q_2,\cdots,q_n\}$,则 MAE 可由下式计算：

$$MAE=\frac{\sum_{i=1}^{N}|p_i-q_i|}{N} \qquad ③$$

因为样本量比较大，只需随机将手机数据集按 4：1 的比例分训练集和测试集即可，准确性影响不大，本文使用 Clementine12.0 “分区”节点按 4:1 进行训练集和测试集分区。对测试集生成各地区与所有故障的评分反余弦矩阵 A，对训练集生成各地区与所有故障的预测评分反余弦矩阵 B。根据公式③计算得到平均绝对偏差（MAE），MAE 越小代表推荐质量越高。由于评分矩阵的生成是根据数据集中记录条数，因此预测评分矩阵与实际评分矩阵在数值上存在差异，不可能得到 MAE=0。因此为了能显示出 MAE 的合理性，分别计算 A 和 B 的平均值$\bar{A}$和$\bar{B}$，通过比较 MAE 与$\bar{A}$、$\bar{B}$，若 M=|MAE-（$\bar{A}$+$\bar{B}$/4）|趋向于 0，即 M 越小代表推荐质量越高。

(2)推荐评价结果分析

运行程序（见附录 8），可算得 M=0.0732，M 接近于 0，因此由该推荐算法产生的推荐结果是比较理想的且可接受的。

3. 结论

本次数据挖掘目的是在维修记录信息进行归一化，区号提取等预处理的基础上，提取手机数据，采用数据挖掘技术，对产品型号、地区、市场级别与反映问题描述（故障）进行关联规则挖掘，基于协同过滤的个性推荐算法的手机数据集分析，从而推出备件的储备需求预测。

（1）由于原始的维修数据存在大量的噪声，后续工作的进行时十分不利的。可以通过其他属性来去除这些噪声数据的，如根据产品型号来填充价格（可认为同一牌子的同一型号的手机价格一样），但分析发现即使是同一牌子的同一型号的手机的价格也相差很远，因此很难替补。本文根据研究目的只选取手机数据集其中的五个属性进行分析，对于缺失数据只是进行简单的删除，因样本量大，故删除掉的少量数据对结果影响不大。

（2）通过利用基于协同过滤的个性推荐算法对手机数据集分析，可知在地理位置上相近的地区具有一定的关联性；不同的地区不同的手机型号常见的手机故障主要是开机故障，触屏故障，按键故障和通话故障，从而可得出备件的储备需求预测；在不同级别的市场购买的手机,，其经常出现故障的手机的手机型号都是 T818，T92，EG906，T912 和 U8，故可为消费者提供购买意见。

(3) 虽然用 Apriori 算法和 GRI 算法得到的模型的支持度或置信度不是很高，但是初步的结果跟用推荐算法得到的结果又一定的相似性，所以 Apriori 算法和 GRI 算法得到的结果可以算是本文粗略的结果，而推荐算法得到的是比较精确的结果。

4. 参考文献

[1] 薛薇.陈欢歌 Clementine 数据挖掘方法及应用 电子工业出版社

[2] 胡洁.张珂珩 数据挖掘在设备状态预测中的应用浅析 江苏瑞中数据股份有限公司

[3] 李卫斌 数据仓库和数据挖掘在航空维修信息分析中应用研究 2010.05

[4] 周张兰 基于协同过滤的个性化推荐算法研究 2009.06

[5] 刘枚莲.刘同存.李小龙 基于用户兴趣特征提取的推荐算法研究 1001-3695(2011)-05-1664-04

[6] 彭石 基于用户兴趣和项目特性的协同过滤推荐算法研究 2012.06

[7] 陶俊.张宁 基于用户兴趣分类的协同过滤推荐算法 2011

5. 附件

附录 1 手机故障及代号

手机故障及代号					
故障类型	故障代号	故障现象	故障类型	故障代号	故障现象
1.开机故障	A1	不开机	10.拍照故障	H1	无拍照
	A2	开机死机		H2	拍照花屏

	A3	开机有电流声		H3	拍照白屏
	A4	开机电流大		H4	拍照黑屏
	A5	开机电流小		H5	拍照彩屏
	A6	开机无声		H6	拍照倒屏
	A7	开机掉电		H7	拍照有条纹
	A8	开机自振		H8	拍照屏闪
2.LCD 显示故障	B1	LDC 显示白屏		H9	拍照阴影
	B2	LCD 显示黑屏		H10	拍照颜色失真
	B3	LCD 显示花屏		H11	拍照死机
	B4	LCD 破屏	11.触屏故障	I1	触屏无效
	B5	LCD 有蓝点		I2	触屏错乱
	B6	LCD 有黑点		I3	触屏难校准
	B7	LCD 有条纹	12.振动故障	J1	无振动
	B8	LCD 有阴影		J2	无振动 INT
	B9	LCD 显示错乱		J3	振动杂音
	B10	LCD 图象反		J4	振动弱
	B11	LCD 无显示		J5	振动强
3.按键故障	C1	按键无效（全部）	13.MP3、收音故障	K1	不识 T 卡
	C2	数字键无效		K2	MP3 声音小
	C3	功能键无效		K3	MP3 杂音
	C4	侧键无效		K4	MP3 播音死机
	C5	拍照键无效		K5	无收音
	C6	导航键无效		K6	收音搜不到台
	C7	功能键手感不良		K7	MP3 音断续
	C8	侧键手感不良		K8	收音死机
	C9	拍照键手感不良		K9	收音重启
	C10	数字键手感不良	14.喇叭故障	L1	喇叭声音小
	C11	游戏键手感不良		L2	喇叭单边有声音
	C12	数字键丝印不良		L3	喇叭有杂音
	C13	功能键丝印不良		L4	插耳机无效
4.通话故障	D1	无发话		L5	耳机单边发音
	D2	无受话		L6	耳机有杂音
	D3	通话声音小		L7	耳机难插难取
	D4	通话有电流声		L8	喇叭无声
	D5	发话声音杂	15.充电故障	M1	无充电
	D6	受话声音杂		M2	充电 INT
	D7	发话 INT		M3	充电器难取难装
	D8	受话 INT		M4	充电死机
5.网络故障	E1	无网络	16.GPRS 故障	O1	搜不到卫星
	E2	无服务		O2	GPS 打不开
	E3	无信号		O3	GPS 信号弱
	E4	信号弱	17.外观故障	P1	离壳
	E5	无法连接		P2	漏螺丝
	E6	功率低		P3	螺丝滑牙
6.灯故障	F1	指示灯不亮		P4	螺丝生锈
	F2	指示灯暗		P5	螺丝未打紧
	F3	指示灯颜色错		P6	缝隙大
	F4	跑马灯不亮		P7	面壳丝印错
	F5	跑马灯暗		P8	按键丝印错

	F6	跑马灯颜色错		P9	面壳有色差
7. 蓝牙故障	G1	蓝牙不能激活		P10	面壳赃污
	G2	蓝牙不能搜索		P11	面壳有暗斑
	G3	蓝牙不能关闭		P12	面壳划伤
	G4	蓝牙测试死机		P13	面壳变形
	G5	蓝牙不能配对		P14	面壳掉漆
	G6			P15	面壳有断差
8. 不读卡		不吃卡、不识卡、不读卡			不能安装软件
9. 电池故障		电池待机时间短			程序错乱
		电池不耐用	18. 其他故障		短信发不出去
		电池充不进电			

附录 2 手机型号数值化程序

```
%
%手机型号数值化
%
function Area()
clear all;clc;close all;
AREA=[];
[~,~,AREA]=xlsread('filename','Sheet1');
AREAsize=size(AREA,1); %数据的记录条数
type=AREA(2:AREAsize,3);   %手机型号位于数据集中的第 3 列
area_type2=unique(type);
area_typesize2=size(area_type2,1);
area_type_type=subs(type,area_type2',{1:area_typesize2});
%将手机型号转换成数值型
xlswrite('Typedouble.xls',Typedouble); %保存矩阵
end
```

附录 3

地区对应数值化的数值表

地区	记录数量	市	省份
28	29	成都市	西川省
002	63	不存在	
006	6	不存在	
010	17467	北京	

020	26984	广州	广东省
021	15312	上海	
022	13084	天津	
023	20880	重庆	西川省
024	9934	沈阳	辽宁省
025	27469	南京	江苏省
027	14903	武汉	湖北省
028	22073	成都市	西川省
029	21198	西安	陕西省
041	449	大连	辽宁省
059	17	福州	福建省
311	21119	石家庄	河北省
315	4401	唐山	河北省
351	16052	太原	山西省
371	26027	郑州	河南省
395	11096	漯河	河南省
411	13364	大连	辽宁省
431	6924	长春	吉林省
451	6045	哈尔滨	黑龙江省
453	1935	牡丹江	黑龙江省
471	5055	呼和浩特	内蒙古
51-	521	韩国	
510	15197	无锡	江苏省
513	6743	南通	江苏省
516	14057	徐州	江苏省
531	26707	济南	山东省
532	65271	青岛	山东省
533	7091	溜博	山东省
535	5142	烟台	山东省
536	4861	潍坊	山东省
537	7531	济宁	山东省
539	23291	临济	山东省
551	25256	合肥	安徽省

555	20	马鞍山	安徽省
571	23722	杭州	浙江省
574	4561	宁波	浙江省
577	13681	温州	浙江省
591	15590	福州	福建省
592	2904	厦门	福建省
710	4476	襄樊	湖北省
731	22191	长沙	湖南省
754	1793	汕头	广东省
755	21739	深圳	广东省
771	16739	南宁	广西省
791	11329	南昌	江西省
797	2794	章州	江西省
817	968	南充	西川省
830	3163	泸州	西川省
833	875	乐山	西川省
851	6644	贵阳	贵州省
871	5206	昆明	云南省
898	8624	海口	海南省
931	6509	兰州	甘肃省
951	47	银川	宁夏回族自治区
991	5397	乌鲁木齐	新疆
空白	2887		
总计	685413		

表 2

附录 4

产品型号对应数值化的数值表

产品型号	数值化	产品型号	数值化	产品型号	数值化	产品型号	数值化	产品型号	数值化
C100	1	E860	21	EG929	41	T830	61	U820	81
C118	2	E87	22	EG939	42	T860	62	U850	82
C210	3	E89	23	EG950	43	T89	63	U850A	83

C268	4	E910	24	EG958	44	T9	64	U860	84
C278	5	E912	25	EG968	45	T908	65	U909	85
E3	6	E912S	26	EG970	46	T909	66	U912	86
E300	7	E920	27	ET919	47	T912	67	U930	87
E310	8	E926	28	LED32H310	48	T92	68	U939	88
E316	9	E930	29	N52	49	T928	69	U950	89
E316B	10	E956	30	N56	50	T929	70	U956	90
E350	11	E956Q	31	S17	51	T930	71	U958	91
E500	12	E968	32	T11	52	T950	72	U960Q	92
E520	13	EG59	33	T30	53	T958	73	U970	93
E6	14	EG617	34	T35	54	T959	74		
E79	15	EG870	35	T39	55	T96	75		
E8	16	EG876	36	T51	56	T960	76		
E820	17	EG900	37	T80	57	T970	77		
E830	18	EG901	38	T81	58	T980	78		
E839	19	EG906	39	T818	59	TG88	79		
E86	20	EG909	40	T820	60	U8	80		

表 3

附录 5 地图及对应区号

附录 6　地区与故障的评分图

```
%
%地区与故障的评分图
function area_pro()
 clear all;clc;close all;
 A_P=xlsread(' holibour.xlsx','area_problem');
 [line,col]=size(A_P);
 x=1:col-1;
 y=A_P(:,1);
   plot(x,[A_P(1,2:col)',A_P(2,2:col)',A_P(3,2:col)',A_P(4,2:col)',...
        A_P(5,2:col)',A_P(6,2:col)',A_P(7,2:col)',A_P(8,2:col)',...
        A_P(9,2:col)',A_P(10,2:col)',A_P(11,2:col)',A_P(12,2:col)'...
        A_P(13,2:col)',A_P(14,2:col)',A_P(15,2:col)',A_P(16,2:col)',...
        A_P(17,2:col)',A_P(18,2:col)',A_P(19,2:col)',A_P(20,2:col)'...
        A_P(21,2:col)',A_P(22,2:col)',A_P(23,2:col)',A_P(24,2:col)',...
```

```
        A_P(25,2:col)',A_P(26,2:col)',A_P(27,2:col)',A_P(28,2:col)',...
        A_P(29,2:col)',A_P(30,2:col)'])
    title('地区与故障');
 end
```

附录 7 推荐算法的运行代码

```
%%
%推荐邻居矩阵
%
%Init 最原始地区与故障的概率矩阵
%Area_area 地区与地区的 cos 矩阵
%Problem 故障的描述
%%
%function result=tuijian()
clear all;clc;
[Init Area_area Problem]=CCOS();%维修设备的测试
neighbour=[];
DIST=0.7;
[line,col]=size(Init);
feel=Init(:,2:col); %去掉 Init 矩阵中的第一列（该列的数值代表服务商所在地 ）
[line,col]=size(feel);
feelmean=mean(feel,2) %求每行的平均值
for i=1:line
    for j=1:col
        sum=0;
        sumArea=0;
        for t=1:line
            if t~=i & Area_area(i,t)>=DIST
                sum=Area_area(i,t)*(feel(t,j)-feelmean(t))+sum;
                sumArea=Area_area(i,t)+sumArea;
            end
        end
        neighbour(i,j)=feelmean(i)+sum/sumArea;
```

```
    end
end
[vals,index]= sort(neighbour,2,'descend') ;   %按行从大到小排列
%vals 排序后的距离
%index 之前的位置
F=Init(:,1);
for i=1:line
    result(i,:)=[F(i,1) Problem(index(i,:),1)'];%邻居矩阵
end
index=find(vals<=0);
result(index+line)=0;
%c=acos(neighbour);
%xlswrite('训练样本集.xlsx',c,'训练样本集'); %保存矩阵
xlswrite('holibour.xlsx',result,'推荐故障（cos>=0.8 ');
xlswrite('holibour.xlsx',vals,'推荐故障 cos 值'); %保存矩阵
%%
% %相近的邻居
% %余弦相似性
function [Feel holibour Problem]=CCOS()
clear all;clc
  [Feel Problem]=areaproblem();
 [line,col]=size(Feel);
feel=Feel(:,2:col); %去掉 Feel 矩阵中的第一列（该列的数值代表服务商所在地 ）
[line,col]=size(feel);
feelst=feel';
for i=1:line
    NORM(i)=norm(feel(i,:));
end
holibour=(feel*feelst)./(NORM'*NORM);   %余弦相似性计算公式
 [vals,index]= sort(acos(holibour),2);    %按行小到大排列
%vals 排序后的距离
%index 之前的位置
F=Feel(:,1);
for i=1:line
```

```
    result(i,:)=[F(i,1) F(index(i,:),1)'];%邻居矩阵
end
result1=result;
result3=result;
index=find(vals>0.3);
result(index+line)=0;
index=find(vals>0.2);
result3(index+line)=0;
xlswrite('holibour.xlsx',result1,'邻居（不做处理）'); %保存矩阵
xlswrite('holibour.xlsx',result3,'邻居（cos>=0.8）');
xlswrite('holibour.xlsx',result,'邻居（cos>=0.7）');
xlswrite('holibour.xlsx',vals,'邻居 cos 值'); %保存矩阵
xlswrite('holibour.xlsx',holibour,'Area_area'); %保存矩阵

%%
%兴趣矩阵的生成
% % %%
 function [area_problem area_type2]=areaproblem()
clear all;clc;close all;
global usetime market Area type  problem
area_problem=[];
AREA=[];
[~,~,AREA]=xlsread('area.xlsx','Sheet1');
%[~,~,AREA]=xlsread('训练样本和测试样本.xlsx','测试样本集');%
% [~,~,AREA]=xlsread('训练样本和测试样本.xlsx','训练样本集');
AREAsize=size(AREA,1); %数据的记录条数
for i=1:AREAsize-1
usetime(i,1)=AREA{i+1,1}; %使用时长  数值型
market (i,1)=AREA{i+1,2}; %市场级别  数值型
Area (i,1)=AREA{i+1,4}; %服务商所在地   数值型
type (i,1)=AREA{i+1,6}; %手机型号数值型
problem (i,1)=AREA{i+1,7}; %反映问题描述  数值型
end
area_type1=unique(Area);
area_type2=unique(problem);
```

```
area_typesize1=size(area_type1,1);
area_typesize2=size(area_type2,1);
area_problem=zeros(area_typesize1,area_typesize2+1);  %服务商所在地 、型号相关矩阵
weixiutime=25-usetime;
area_problem(:,1)=area_type1;
for a=1:area_typesize1
    for p=2:area_typesize2+1
        sum=0;
        for i=1:AREAsize-1
            if Area(i)==area_type1(a) & problem(i)==area_type2(p-1)
               sum=sum+weixiutime(i);
                end
        end
        area_problem(a,p)=sum;
    end
end
 [line col]=size(area_problem);
%c=acos(area_problem(:,2:col));
%xlswrite('测试样本集.xlsx',c,'测试样本集');
% xlswrite('训练样本集.xlsx',area_problem,'area_problem');
```

附录 8

```
%%
%%%误差验证
%fact  实际矩阵
%result 预测矩阵
 function WEA()
 clear all;clc;
fact=xlsread('测试样本集.xlsx','测试样本集');
result=xlsread('训练样本集.xlsx','训练样本集');
[line,col]=size(fact);
A=mean(mean(fact));
B=mean(mean(result));
M=0;
```

```
for i=1:line
    sum=0;
    for j=1:col
        sum=abs(fact(i,j)-result(i,j))+sum;
    end
    M=sum/col+M;
end
M=abs(M/line-A-B/4)
```