应用聚类分析观点多样性

穆罕默德·马赫迪·哈桑
瑞典卡尔斯塔德大学计算机科学系
mohammad.hassan@kau.se

09Martin布鲁姆
瑞典卡尔斯塔德大学计算机科学系
Martin.Blom@kau.se

摘要

在实证软件工程研究中，越来越多地使用问卷和调查来收集从业人员的信息。通常，这样的数据会基于整体的描述性统计数据进行分析。尽管这可以捕捉到总体趋势，但存在丢失不同(少数)子群体意见的风险。在这里，我们建议使用聚类来对受访者进行细分，以便实现更详细的分析。我们的发现表明，它可以更好地了解调查人群和参与者的意见。这种划分方法可以更精确地显示不同群体之间意见差异的程度。这种方法也为少数群体提供了表达意见的机会。通过这一过程，还可能获得重大的新发现。在我们关于工业中测试和需求活动的状态的示例研究中，我们发现了几个显著的组，它们从总体结论中显示出显著的意见差异。

关键字

实证调查，聚类，数据挖掘，划分，分组，多样性，少数，专家意见。

1. 介绍

在软件开发的过程中，会产生许多不同形式的数据。传统的数据形式如下:[1]- 1)代码库，2)执行轨迹，3)历史代码变更，4)Bug数据库等。

在最近，大量的投资被投入到软件过程自动化中，因为它可以降低开发成本，提高产品质量。过程自动化不仅可以大量产生一些传统形式的数据，还可以提供存储和提取新形式数据的机会。其他一些形式的软件工程数据可以描述如下:

*EASE'15,2015*年4月27 - 29日，中国南京版权2015 ACM
978-1-4503-3350-4/15/04···$15.00。http://dx.doi.org/10.
1145/2745802.27458

测试用例——在自动化测试过程中大量使用。测试用例可以手动生成，也可以通过自动化过程生成。

系统构建跟踪——组件构建及其集成过程已经高度自动化，因此可以更容易地跟踪。

· 团队和个人数据——商业工具可以收集和跟踪开发人员(以及团队)的工作和工作模式。

· 开发过程数据——也存在收集完整开发过程数据的工具。

在本文中，我们分析了一个传统的数据源，“意见调查”，它通常不被考虑用于数据挖掘(DM)。我们的研究表明，这种形式的数据可能有一些潜在的模式，可以通过聚类过程来提取。它可能揭示新的信息，并吸引人们对其他视角的注意。

从软件开发从业者那里收集调查数据进行统计分析是软件工程研究[2]的关键领域之一。近年来，在线设施和工具使得频繁地收集调查意见变得更加容易，因此在大多数软件组织中都存在着相当数量的调查数据。

我们注意到，大多数的调查分析是使用传统的统计方法和措施(如均值，中位数，方差和一些数据分析测试)来进行他们的调查结果[3]。总的来说，他们将整个调查人群视为一个单独的组，并使用一些抽样技术来提取[4]品种。在某些情况下，种群也根据一些背景信息被划分为子组[3]。这并不能很好地揭示观点多样性，因为相似的观点可能存在于不同的群体中，而同一群体中的人可能有不同的观点。

在这项研究中，我们将聚类技术应用于以类别或数字形式收集的民意调查数据。没有任何感知偏差的聚类将人群划分为不同的子人群簇，这些子人群在某种程度上具有相似的意见。我们观察到使用这种方法有一些好处和机会，例如:

· 它可以减少对分组的操纵，因为它根据他们的意见来分组。此外，如果有必要，在应用聚类时，背景信息也可以与意见结合在一起。

· 它可以更精确地展示人群中的意见差异。统计方差[3]只能显示总体的一致或不一致，而分组通过DM(聚类)可以显示各组的差异和组内的一致和不一致。

· 它可以识别少数群体，否则将无法识别。在大多数情况下，由于结果以一种更集中的方式呈现，少数群体失去了他们的声音。

· 意见差异和背景信息可能揭示出具有明显特征的群体，这可能会导致产生有效的假设。因此，可以设计进一步的研究来调查这些群体和相关的假设。

· 在某些情况下，不同观点之间的特定形式的相关性仅在集群内可见，并且在集群形成之前并不明显。

为了调查聚类方法，我们使用了由瑞典咨询公司进行的一项调查。根据调查，他们得出了一些总体结论，如测试的状态是令人满意的，但不是需求过程。在我们的研究调查中应用聚类后，我们发现调查人群中有一些相当大的群体，他们的意见在强度和方向上都与总体结论不同。

我们组织我们的论文如下:第二节包含相关工作，在第三节我们试图回答一些可能出现在聚类过程中的问题，在第四节我们描述了样本调查的背景，我们在我们的研究中使用，在第五节我们描述了在类别数据上应用聚类的过程，第六节致力于数字数据聚类，在第七节我们进一步讨论我们的方法。最后，我们在第8节中总结。

2. 相关工作

根据我们对“利用数据挖掘技术进行专家意见调查数据分析”的调研，软件工程在这方面的研究还比较欠缺。我们发现，市场营销[5]和商业相关领域在调查数据[6]上已经做了挖掘研究。这些数据大多来自普通人;由于互联网的繁荣，商业组织通过网络表格、[7]以及传统方法(电话、纸质等)收集客户意见变得更容易了⊠aA⊠Z’的意见，并以数字形式保存它们。他们的关键目标是识别潜在客户和产品，这最终会增加他们的业务和收入。

对从业者的意见调查是软件工程领域的核心研究实践之一。分析这样的调查数据有一些标准指南，基本都是一些基于简单统计方法的理性调查方法。Barbara a . Kitchenham[3] &[9]描述了其中的一些方法，并对使用贝叶斯分析等高级统计方法提出了警告。他们提到⊠AIJBayesian⊠方法通常不用于软件工程研究⊠aA⊠I·，并建议向统计学家寻求帮助。最近John Moses [10]、[11]和[12]提出了一种基于专家意见的利用贝叶斯推理和马尔可夫链蒙特卡洛(MCMC)仿真的软件质量预测模型。在一般情况下

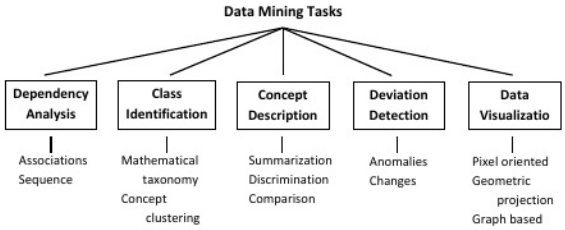


图1:数据挖掘任务分类[15]

描述性统计技术以及假设检验被用于分析调查意见[13]。

在调查研究领域，我们可以观察到一种使用高级统计模型的不安。对于基于高级数学和统计模型[14]的机器学习技术来说，情况也是如此。由于这些调查的数据量很小，研究人员也不鼓励使用DM技术进行分析。在后续章节中，我们将展示在调查数据上应用DM有很好的机会发现不同的视角，这些视角可能会被忽视，而使用传统的理性和更简单的统计分析可能无法发现。

在描述我们的主要研究之前，我们简要概述一下在我们的分析过程中使用的相关数据挖掘任务。根据Shaw et. al. [15]，DM任务可以分为以下组成部分(也见图1)- 1)依赖分析，2)类识别，3)概念描述，4)偏差检测，5)数据可视化。

基本上，我们对样本调查数据进行了群体识别(类识别)和表征(概念描述)。分析过程的细节将在后面章节中描述。

3. 聚类——一些相关问题

我们使用数据挖掘技术对调查数据进行定量分析，然后我们使用聚类产生的结果来发现新的事实或解释一些表明意见多样性的现象。在这个过程中，以下几个问题指导着我们:

我们应该使用哪种工具?

· 我们应该使用哪种聚类算法?

· 我们应该使用多少个问题进行挖掘?

· 有多少个簇?

· 如何单独分析聚类?

在接下来的章节中，我们将尝试回答这些问题。

3.1 我们应该使用哪种工具?

有很多商业以及开源的数据挖掘工具可用[16]。一些挖掘工具提供了丰富的展示工具包，这些工具包可以对理解聚类结果有用。我们使用了开源的数据挖掘工具Weka (http://www.cs.waikato.ac.nz/ml/weka/)。它包含了突出的聚类算法，并提供了一个简单的

可视化界面。它还支持导出聚类后的结果，可以导入(次要处理)到其他统计工具进行进一步分析。

3.2 我们应该使用哪种聚类算法?

基本上有两种聚类算法，分区(或K-clustering)和层次聚类[17]。根据分析的目标，首先应该选择一种类型的聚类范式，然后选择一种具体的算法进行进一步处理。选择合适的算法可能会对结果及其解释产生影响。在选择算法时，我们应该考虑几个因素;数据类型(数值型、类别型等)、数据集的大小和数据集的特征(如缺失数据率、异常值的存在和数据点的分布等)。

我们使用期望最大化(Expectation Maximization, EM)[18, 19]算法进行聚类。它是一种聚类分析方法，旨在将n个观测值(数据点-在我们的案例意见记录中)分类为k个聚类(组)，其中每个观测值根据其可能性属于特定的聚类。EM是一种稳健的算法，因为它可以处理意见调查中常见的缺失数据。EM算法基于强大的统计基础，具有高度的可扩展性和线性数据库大小。它可以处理高维数据，并且可以手动设置预期的聚类数目。通过良好的初始化，它可以快速收敛[20]。除此之外，分类和数值数据都可以使用EM进行挖掘，这是一些其他流行的聚类算法所不支持的，如K-means算法[21]。

3.3 我们应该使用多少个问题进行挖掘?

一般来说，每个调查问题都可以被认为是一个可能用于聚类的属性。属性选择是数据挖掘中一个具有挑战性的问题[22,23]。我们可以通过只考虑相关属性来降低计算成本。它也有助于以更好的形状[22]提取信息。聚类属性的数量没有特定的限制;相反，它取决于数据集的性质。一些属性选择技术，如包装器子集评估[24]，CFS(基于相关性的特征选择)[25]和PCA(主成分分析)[22]可能在这方面有所帮助。识别研究兴趣也可能在这方面有指导作用。除此之外，由于样本量小，一旦在聚类中使用过多的问题，调查意见数据很容易产生不确定的聚类。在我们的案例中，我们只在聚类过程中使用两到三个问题来划分数据集。

3.4 多少个聚类?

缺乏数据集的先验知识很难预测可能的簇数[26]。在像EM这样的聚类算法中，我们可以设置一个先验数字，以便数据集将相应地进行划分。有一些统计措施，如聚类模型的简单分布[27]-[1]，标准差(用于数值数据)，对数和最大似然[18,28]等，可以帮助我们确定可能的聚类数量。我们将在接下来的章节中通过一个例子更详细地讨论它。

3.5 如何单独分析聚类?

选择合适的指标来呈现集群的特征是很重要的。在大多数情况下，标准的统计指标，如百分比，计数，平均/中位数方差等，都用于此目的。指标的选择可以在观察簇之间差异的同时产生影响。我们使用聚类来划分数据集，以便相似的意见被重新组合成簇。我们期望那些更小的组将更有凝聚力，并将显示出一些意见变化。最初，我们分别调查每个聚类，并使用每个属性的统计指标(即每个调查问题)将它们彼此比较以及与一般人群进行比较。通过这个过程，我们还可以观察到不同属性之间的一些相关性。依赖属性(即用于聚类)与独立属性(即不用于聚类)之间的显著相关性可能表明意见差异的原因。

3.6 我们可以重复聚类过程吗?

如果参数设置以及数据集都是完整的，聚类过程是高度可重复性的。在我们的例子中，我们使用Weka提供的默认参数设置，包括初始随机种子，我们只改变了预期的聚类数值。为了支持研究的复制，我们可以提供参数设置的值[2]。跨不同数据集的复制是可能的，但具有挑战性，因为不同数据集的集群可能是不一致的[29]。

4. 样本调查概述

数据是通过一份由IT教育和咨询公司QETMA[30]分发的在线问卷从许多瑞典IT公司收集的，这是他们评估IT行业实践现状工作的一部分。问卷包括21个问题，从个人和企业数据的背景问题，到与IT和参与者使用的流程相关的各种技术和非技术方面的问题。

调查问卷每年发放一次，本研究基于2010年的调查问卷。153名受访者参与了调查。我们选择包括类别型和数值型数据来评估数据挖掘方法——针对这两种类型的问题。用于聚类的问题列在下面[3]，以及答案和选择该答案的受访者的百分比answer[4]:

1. “以下哪句话最好地描述了你最常用的开发方法?”

[1]In Weka’s website it is mentioned as normal distribution which might be misleading, as the distribution table pro-vided by Weka after EM clustering only provides the distri-bution of instances in different clusters for each category for a particular attribute.

[2]Following parameters is used in Weka aAZs´EM implemen-tation aA S 、1) Maximum number of iterations = 100, 2) Min-imum Standard deviation= 1.0E-6, Seed =100 and Number of clusters =-1.

[3] These questions are related to requirement and testing which are the main focus of Qtema. Initially we also tried with some other questions but did not find interesting pat-terns. For larger datasets there might be a need for a sys-tematic attribute reduction process to select clustering at-tributes.

[4]These have been translated from the Swedish original by the authors

(a)传统的、计划驱动的流程，需求之后(按顺序)是设计、实现和测试，30%

(b)需求、实施和测试工作都在短迭代中完成的敏捷过程，14%

(c)有意识地将传统的、计划驱动的流程与使用短迭代的敏捷流程相结合，52%

(d)其他，4%

2. “你的公司/单位是否有一个有效的组织和流程来满足需求?(a)非常低的比例，10%

(b)低程度，45%

(c)高程度，35%

(d)非常高，8%

3. “你的公司/部门在测试/验证/确认方面有一个有效的组织和程序吗?”

(a)非常低的比例，3%

(b)低一点，26%

(c)高程度，56%

(d)非常高的程度，15%

4. “你们在需求上花了多少时间(单位:%)?”[5] (a) <20%， 60%

(b) >30%， 3%

(c)均值，15%

5. “你们在测试、验证和确认上花了多少时间(单位:%)?”

(a) <20%， 31%

(b) >30%， 12%

(c)均值，22%

从21个可能的问题中选择这些特定问题是基于QTEMA的概况，该概况侧重于需求和测试，以及这些领域与所使用的开发方法之间可能的相关性。问题1至问题3用于类别数据聚类，其余两个用于数值数据聚类。为简单起见，我们将使用Traditional, Ag-ile和Blend分别作为选项a)， b)和c)的问题1的替换答案。

我们还纳入了研究中我们认为有趣的其他一些问题，这些问题与受访者的工作习惯有关。下面列出了调查中使用的这些问题和备选答案，以及它们的相关统计数据。

6. “你目前的工作经验有多少?”

(a) <1年，6%
(b) 1-3年，18%
(c) 3年以上，76%
(d) 5年以上，56%
(e) 10年以上，33%

7. “你具备专业工作所需的能力吗?”

(a)非常低的程度，0%

(b)低一点，4%

(c)高程度，56%
(d)非常高，38%

8. “以下哪项最好地描述了你在一个典型的开发项目中什么时候会见客户?”’ (a)从来没有，7%

(b)一开始，5%

(c)最后，4%

(d)开头和结尾，21%
(e)在整个项目中连续的和多次的情况下，54%

(f)在整个项目中每天进行，8%

9. “你在实现/编码上花了多少时间(单位:%)?”

(a) <20%， 79%

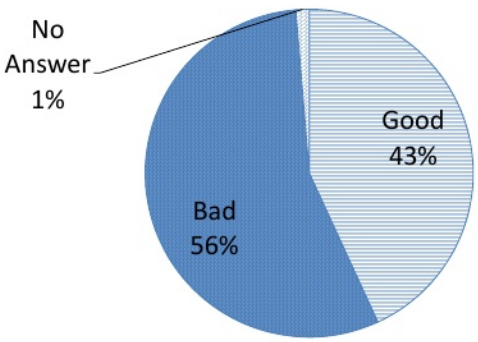
(b)均值，34%

10. “你在设计/分析/建模上花了多少时间(%)?”

(a) <20%， 16%

(b)平均值，15%

图2:总体需求状态



根据调查，Qtema建议，需求相关激活的状态很差，但总体测试激活是令人满意的[6]。我们分别在图2和图3中给出了他们的总体结论。

[5]这个和下面的问题已经从一个关于时间消耗的复合问题中分离出来，并进行了重写，以增加可读性。最初的问题是——“你在以下活动上花了多少时间(百分比)?”(对于每个活动，你可以声明一个0-100之间的数字，但所有活动的总和应该是100)”。由于受访者可以在这些问卷中自由输入数字，所以我们做了分类(基于切割点)。除了呈现受访者的百分比，我们还提供平均值。

[6]这里“到很低的程度”和“到很低的程度”被认为是不好的，而“到非常高的程度”和“到很高的程度”被认为是好的

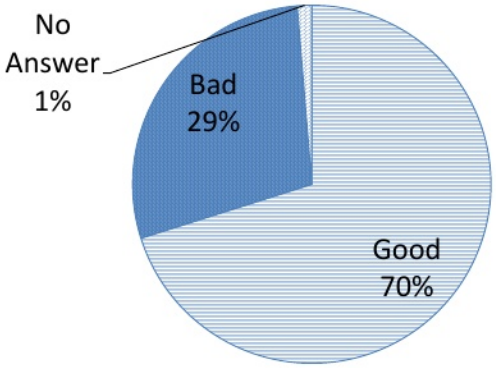


图3:整体测试状态

5. 聚类分类数据

表1:调查问题1 - 3的数据透视表。“TVL”:非常低的程度，“TL”:非常低的程度，“TH”:非常高的程度，“TVH”:非常高的程度，

	SQ1 (Process)	SQ2 (Requirement)			
		TVL	TL	TH	TVH
SQ3 (Testing)	Traditional				
	TVL				
	TL	5	8	2	
	TH	1	15	10	
	TVH		1	1	3
	Agile				
	TVL	2			
	TL	1	2	2	
	TH	1	4	6	
	TVH			1	2
	Blend				
	TVL			1	
	TL	2	12	4	1
	TH	4	22	18	1
	TVH		3	8	3

聚类分类数据有一种内在的支持，由于选项有限，人们可以根据所有类别的组合来预测最大可能的组。如果用于分组的列数(每列包含单个问题的意见)是N，每列分别有m_1，$m_2 \cdots m_n$类别数那么可能分组的最大数目可以简单地计算如下式1所示:

$$Maxclusters = m_1 * m_2 * m_3 \ldots * m_n \quad (1)$$

在表1中，我们显示了基于调查问题1到3的枢轴表，它们代表绝对分类。在48个可能的群体中，只有5个群体的人口数量大于或等于10，这表明整个群体被分成了不重要的群体。因此，在现实中，我们可能会发现，由于绝对分类的原因，这些可能的群体中有很大一部分是空的或不重要的，只有少数几个有意义的群体存在。

自然，对于分类数据来说，可能的组的数量

表2:簇分布，簇集个数为1

SQ	Answers	Cluster 1
1	Traditional	47
	Agile	22
	Blend	80
	Others	8
2	Very low	17
	Low	70
	High	55
	Very High	13
	No Answer	3
3	Very low	5
	Low	41
	High	86
	Very High	24
	No Answer	3

随着问题数量的增加而增加。在一些调查中，数值也被用来表达观点(如规模问题)，这极大地增加了可能选项的组合。因此，在大多数情况下，单独观察和分析所有可能的群体是不可能的。此外，在更复杂的情况下，我们可能会发现，在人口的某些部分，某些类别是过时的，但在其他部分，它们不是。在这种情况下，绝对的分类可能会误导分析过程。另一方面，我们的兴趣不是所有群体，而是有不同意见的一些群体。有许多聚类算法可以通过识别重要的群体来更容易地处理这种情况。

在接下来的章节中，我们将描述在我们的抽样调查中应用聚类的过程。

5.1 初始步骤

我们使用调查问题(SQ) 1到3进行分类聚类(参见第4节)。最初我们将预期的聚类设置为1，在表2中我们显示了Weka的EM实现提供的聚类结果[7]。结果显然表明，对于每个问题，都有显著的意见差异(如在SQ2中，对于每个可能的答案，都有超过10个人)。在接下来的步骤中，我们将尝试将总体人群划分为不同的部分，以便在每个部分中多样性减少，类似类型的意见(在所有三个问题中)鲜明地显示自己。

5.2 分区

在Weka的EM实现中，分区可以通过两种方式进行:1)使用默认设置，其中预期的聚类数设置为-1，这表明EM将自动确定数量[8];2)设置预期的聚类数并观察结果。

在表3中，我们显示了默认设置的结果，该设置产生了四个簇。在这里，我们使用两个索引来表示一个集群，就像在C(-1,1)中，第一个索引建议预期

[7]Note default value for each category is preset to 1(probably to support logarithmic operation), so the actual distribution is bit different, ex. Like 47, 22, 80 and 8 in the third column of table 2 for question one, actually the numbers of instances are 46, 21, 79, and 8 which in total is 153, the number of opinion records we started with.

[8]Using some form of cross validation technique[27]

集群设置是默认的-1，第二个索引表示一个特定的集群。我们在表3到表8中遵循相同的格式来表示聚类结果。

表3:簇分布，簇集数量为默认值-1

SQ SQ	Answers Answers	C (-1,1)	C (-1,2)	C (-1,3)	C (-1,4)
1	Traditional	9.32	3.60	35.71	1.38
	Agile	9.22	2.68	9.64	3.46
	Blend	32.67	3.032	45.92	1.37
	Others	2.36	4.05	2.27	2.32
2	Very low	2.42	1.29	12.99	3.31
	Low	8.58	1.78	60.64	2.01
	High	37.04	1.63	17.95	1.38
	Very High	5.36	7.46	1.88	1.30
	No Answer	1.18	2.20	1.09	1.53
3	Very low	1.98	1.34	1.38	3.30
	Low	5.45	1.46	35.26	1.83
	High	31.29	2.72	53.19	1.80
	Very High	14.84	7.16	3.71	1.29
	No Answer	1.01	1.67	1.00	1.31

在下一步中，我们开始使用手动设置的预期簇数(为了方便起见，我们将其表示为N)来研究聚类过程。首先，我们将N设置为2。在表4中，我们展示了结果。集群C(2,2)提出了一个新的观点，因为集群中的大多数人对需求过程(SQ2)很有信心。另一方面，集群C(2,1)看起来不连贯，因为同一问题在同一个集群中有多个主导选择。因此我们继续增加N的值并分析结果。

5.3 识别有趣的聚类

我们的目标是观察人口群体中的不同意见。因此，在获得每个设置的聚类结果后，我们试图定位那些大小重要的聚类，并显示某种形式的替代结论。在不同的集群中，意见差异由每个问题中意见类别之间的人口分布来表示。我们检查每个问题(用于聚类)分别为每个簇观察差异

表4:聚类分布，聚类集合的数量

2

SQ	Answers	C(2,1)	C(2,2)
1	Traditional	41.94	6.06
	Agile	17.65	5.35
	Blend	64.04	16.96
	Others	3.17	5.83
2	Very low	16.30	1.70
	Low	66.61	4.39
	High	41.84	14.16
	Very High	1.90	12.10
	No Answer	1.14	2.86
3	Very low	3.78	2.22
	Low	39.43	2.57
	High	78.17	8.83
	Very High	5.40	19.60
	No Answer	1.01	1.99

表5:集群分布，有趣的集群贯穿整个挖掘过程

SQ SQ	Answers Answers	C (5,1)	C (6,6)	C (6,2)	C (6,4)	C (7,7)
1	Traditional	9.31	3.99	1.93	15.38	11.36
	Agile	9.17	4.66	5.46	1.70	1.66
	Blend	32.21	4.16	19.33	1.78	1.70
	Others	2.37	2.39	1.90	1.41	1.50
2	Very low	2.56	2.54	3.93	6.29	1.41
	Low	8.97	1.38	15.87	9.70	1.68
	High	37.61	1.9	6.78	3.01	11.59
	Very High	3.73	9.13	1.97	1.19	1.30
	No Answer	1.20	1.22	1.08	1.08	1.23
3	Very low	1.98	2.41	1.46	1.18	1.07
	Low	5.53	1.35	23.49	15.43	1.72
	High	32.53	1.47	1.88	1.83	11.30
	Very High	13.02	9.74	1.73	1.77	2.02
	No Answer	1.01	1.20	1.08	1.07	1.11
Size		**50**	**12**	**25**	**15**	**13**
Group		**1**	**2**	**3**	**4**	**5**

过。如表3所示，四个集群中的两个在大小上不显著，C(-1,2)和C(-1,4)分别是人口的6%(9人)和2%(3人)。另一方面，C(-1,1)和C(-1,3)的规模分别为28%(43人/子)和64%(98人)的人口。在这里，聚类C(-1,3)与一般结论一致(对需求过程的置信度低)，而聚类C(-1,1)与一般人群相比有不同的意见，因为绝大多数人在SQ2(对需求过程的置信度更高)中做出了积极的回应。

在定位了有趣的簇之后，我们将它们从群体的其余部分中分离出来，并单独分析它们。相同的集群可以出现在不同的设置中，随着N的增加，这些集群要么重新出现，要么解体为更小的集群。如果我们发现一组在N的不同设置中出现的集群，但在用于聚类的所有属性中表现出共同的特征，那么我们将它们视为同一组的不同表示。在表5中，我们显示了这些组的相关分布统计数据。在这里，我们根据N的设置放置结果，其中每个组的代表性集群的规模最大。

5.4 停止集群

因为我们的兴趣是定位不同的观点，所以找到一个群体比整体分类更重要。我们继续增加N的值，并寻找感兴趣的聚类，即使整体分类恶化率(见图4中预期聚类的不同设置的对数似然)。在将N的值设置为7后，我们没有发现任何呈现新观点的显著聚类。在图5中，我们展示了在N的不同设置下，不同组的大小变化。

5.5 分析显著组

我们使用第4节中提到的所有调查问题来分析我们通过聚类过程提取的每个组[9]。正如预期的那样，它们在SQ1到SQ3中都显示出了一些区别，因为它们被用于聚类，但它们也

[9]Note we did not find any significant patterns in SQ9 and SQ10 for any groups.

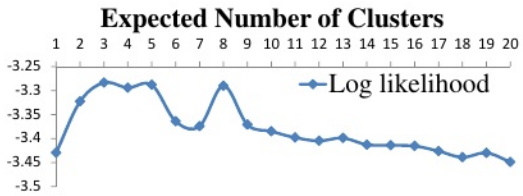


图4:不同预期聚类数设置的对数似然

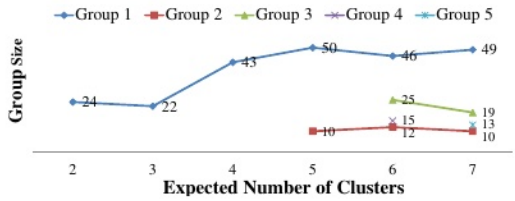


图5:表5中显示的不同组的大小变化

在剩下的问题上显示一些有趣的差异。在下一节中，我们将介绍每一组，并描述它们的一些独特特征。

1. 第1组:这是持有不同意见的最大的组，其中大多数人对需求和测试过程都很有信心。这个群体的一些独特属性:

   (a) 94%表示对需求过程的信心，98%表示对测试过程的信心(即在SQ2或SQ3中选择c或d)。注意，在一般人群中，43%和70%分别表示对需求和测试过程的置信度。

   (b) 76%表示与客户进行某种形式的频繁互动(在SQ7中选择e或f)，而在一般人群中，这一比例为62%。

2. 第二组:这是一个小团队，所有成员对需求和测试过程都非常有信心。这个群体的一些独特属性:

   (a)除了一个没有回答的成员，在SQ2或SQ3中都选择了d。

   (b)他们经验丰富;12个成员中有7个(58%)有超过10年的经验，在一般人群中这一比例为33% (SQ7)。

   (c)他们对自己的专业能力非常自信;12个成员中有8个(67%)在SQ8中选择了d，在一般人群中这一比例为33%。

3.第3组:这个组不仅对需求过程信心不足，而且对测试过程也信心不足。这个群体的一些独特属性:

   (a) 100%的成员在SQ3中选择了b(表明对测试过程的信心更少)，在一般人群中只有29%(结合a和b)。

   (b)在SQ5中，他们有一个平均值(16%)，比一般人群(22%)低得多。这可能表明他们较少参与测试过程。

4. 第4组:这与第3组类似，但规模更小，由遵循传统开发过程的人组成。这个群体的一些独特属性:

   (a) SQ1中100%的成员选择了a。

   (b) 100%的成员对测试过程表现出更少的信心。

   (c) 60%的成员认为他们不经常与客户互动，这比一般人群高得多(37%的人在SQ8中选择a到d)。

5. 第5组:这一组与第1组一致，但规模更小，并由遵循传统开发过程的人组成。这个群体的一些独特属性:

   (a) SQ1中100%的成员选择了a。

   (b) 100%的成员对需求和测试过程都表现出信心。

   (c)对于SQ5来说，他们的平均值高于一般人群(29% vs 22%)，这可能表明他们更多地参与测试过程。

通过考虑我们省略的其他调查问题，可能会发现更有特色的属性。另一方面，在这篇论文中，我们的主要目的是表明存在强有力的替代意见，并且聚类技术可以揭示它们。

6. 聚类数值数据

在调查中，数值数据用于捕获许多不同形式的信息。有时它被用来收集像大小或价值这样的直接信息，其他时候它被用来通过某种形式的数字尺度收集相对信息，比如在SQ4和SQ5中，信息是以百分比的形式收集的。在这两种形式的数值数据中，我们可以应用聚类技术来识别一些有凝聚力的群体，这些群体可能显示出一些明显的特征。

在接下来的章节中，我们将描述数值聚类的过程。

6.1 初始步骤

我们使用第4节中提到的SQ4和SQ5对数值数据进行实验。这两个问题都与参与者的工作习惯有关。

首先，我们将N的值设置为1，以获取一般人群的相关统计数据。我们将结果显示在表6-[10]的第三列。这两个问题的标准差都非常高，这表明调查参与者的工作习惯可能存在显著差异。

[10]M=Mean and SD=Standard Deviation in 2nd Column in the tables

表6:默认设置下数值型数据的聚类结果

SQ	Stat	In Total Population	C(-1,1)	C(-1,2)
4	M	15.26	11.33	18.70
	SD	8.15	5.44	8.57
5	M	22.37	26.76	18.52
	SD	9.71	11.13	6.07
Size		**153**	**77**	**76**

表7:在N的不同设置下，组1的代表性聚类

SQ	Stat	C (7,4)	C (8,1)	C (9,1)	C (10,1)	C (11,7)
4	M	10	9.72	9.39	12.05	9.07
	SD	8.18	3.58	3.63	4.83	2.09
5	M	17.02	18.9	18.91	16.60	19.99
	SD	5.70	3.33	2.91	2.87	.016
Size		**19**	**29**	**25**	**15**	**16**

通过聚类，我们将尝试揭示出根据工作习惯(在本例中仅为测试和需求)而彼此不同的组。稍后我们将对他们进行单独分析，以观察由于工作习惯而可能出现的意见差异。

6.2 分区

像分类数据一样，Weka也支持聚类数值数据的两种模式:1)聚类的预期数量N的先验设置，2)默认值N为-1，以通过交叉验证获得聚类的数量。与分类聚类相比，聚类过程后的结果呈现是不同的，因为它以每个聚类的平均值和标准差的形式呈现结果。

我们从默认设置开始，它产生两个聚类C(-1,1)和C(-1,2)，如表6所示。两个簇都显示了更高的标准差;由于我们正在寻找有凝聚力的群体，我们继续进一步聚类。

6.3 识别有趣的集群

我们根据均值和标准差识别有趣的簇。我们期望一个有凝聚力的群体在聚类中使用的每个属性上的标准差都要低得多(与一般人群相比)。我们还期望这些群体中的每个人在某些属性中将具有不同的均值，这将使他们与众不同。在确定组之后，我们在N的不同设置下观察和分析组，以选择最能代表组的簇。这里我们主要考虑标准差和大小，使群体足够大，具有可接受的标准差。

在表7中，我们显示了一个组(我们将其表示为组1)，在不同的n设置下，与一般人群相比，在SQ4和SQ5中具有较低但一致的均值。这里C(11,7)考虑到标准偏差是最具凝聚力的，而C(8,1)是最大的，具有可接受的标准偏差。在这里，我们避免给出任何硬规则来选择任何特定的集群，相反，我们的目的是表明我们可以使用聚类来识别这样的群体。在分析阶段，考虑到他们在其他调查问题中的意见，我们将表明两个簇代表一个非常相似的组。

6.4 停止集群

在图6中，我们展示了不同N值下规模为10%人群的集群数量的比较。从图中可以明显看出，具有显著人群的集群数量是稳定的，最初上升到5，然后在N值较高时下降到3。另一方面，最初没有更小的群体，但在它们第一次出现后，它们的数量几乎与N线性。

当我们增加N的值时，我们开始得到较小的集群，N的值很高，得到具有显著人口的新不同组的机会很小。因此，我们可以停止聚类过程。

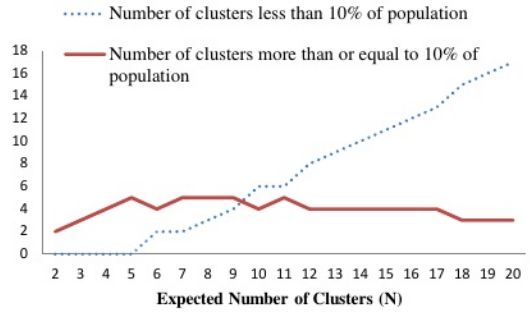


图6:预期聚类数(N) vs包含大于或小于10%人口的聚类数

6.5 分析显著组

我们已经提到，对于n的不同设置，一个组可以用不同的簇来表示。在表7中，我们显示了代表组1的不同簇。首先，我们将分析C(8,1)和C(11,7)分别代表组1的最大和最凝聚的簇。稍后我们将分析两个更有趣的组，如表8所示。

1. 组1:这个组代表明显较少参与需求活动的人(在SQ4中，平均值约为9，与一般人群相比少40%)。他们在测试活动中的参与度也略低。

   (a) C(8,1):这是代表组1的最大集群。在这个集群中，66%的成员对需求过程表现出更少的信心，这高于一般人群(56%)。他们与客户互动的频率也较低(在SQ8中为52% vs 62%)。

   (b) C(11,7):这是代表组1的最具凝聚力(考虑到标准偏差)的集群。在这个集群中，75%的成员对需求过程表现出更少的信心，这比一般人群高得多。他们与客户互动的频率也较低(在SQ8中为50% vs 62%)。

表8:第2和第3组

SQ	Stat	C(9,6)	C(7,2)
4	M	8.61	28.06
	SD	3.95	2.47
5	M	30.67	17.39
	SD	2.50	5.68
Size		24	16
Group		Group 2	Group 3

因此，除了在SQ4和SQ5中具有相似的平均值外，两个集群在意见上共享了我们预期的相似趋势。

2. 第2组:这一组代表了调查参与者中的一部分，与一般人群相比，他们更多地参与测试。与普通人群相比，他们对测试过程显示出明显更高的置信度(在SQ3中为88% vs 70%)。他们在需求过程中也表现出了更高的信心(54% vs 43%)。

3.组3:这个组代表更多参与需求相关工作的调查参与者(在SQ4中，他们的平均值为28，比一般人群高87%)。与一般人群相比，他们对需求过程更满意(在SQ2中为63% vs 43%)。他们也与客户有更高的互动(在SQ8中94% vs 62%)。

7. 讨论

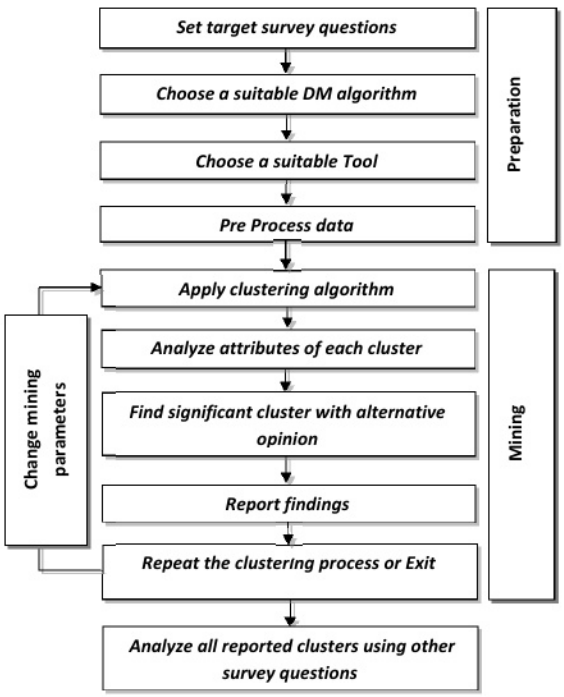


图7:流程示意图

Anderberg表明，通过简单的人类能力从数据集中理解可能的分组是极其困难的。他举了一个例子，将25个观测值可能分组为5个组是巨大的(精确为2,436,684,974,110,751)[31]。因此，即使是一个小的调查，也几乎不可能手动彻底地划分人口并调查他们的特征。但是聚类在这方面可以有所帮助，因为它被用于其他领域来解决类似的问题。

在这项研究中，我们以一种系统的方式应用聚类技术来划分调查人群，然后对显示不同意见的重要群体进行分离和分析。在图7中，我们使用流图展示了这个过程。在前面的章节中，我们已经通过示例讨论了该过程。这里我们讨论一些可能阻碍这一过程的重要因素。

在开始挖掘过程之前，准备数据是一项重要的活动。有些算法不够健壮，无法处理缺失数据，因此需要将空记录剔除，或者用一些有意义的数据填充，以将它们与其他记录区分开来[11]。在目前的研究调查中，我们看到很少有参与者不回答问题(比如只有两名参与者没有回答SQ1)。

在某些情况下，参与者根据自己的理解归纳出非标准信息。就像在这项研究调查中，153名参与者中有4名提供了SQ1中的进程名称，这些名称不在列表中(归入“其他”类别)。在这项研究中，我们没有更改他们提供的数据，但在其他一些情况下，如果此类案例的数量很高，那么修改它们可能会导致更好的聚类。

8. 结论和未来的工作

数据挖掘在基于意见的调查中的应用在软件工程中并不常见;可能调查参与者数量少，可能会阻碍研究人员使用DM作为分析工具。另一方面，在调查人群中可能存在一些较小的群体，他们可能有显著的视角差异，可以导致未来的成功或警告失败。在传统方法中，很难观察到不同群体之间的意见多样性，因为发现它们是具有挑战性的。因此，在大多数分析中，某种形式的整体统计指标被用来获得一种全面的了解，这削弱了意见多样性。在我们的研究中，我们展示了一些常见的聚类工具和技术可以帮助我们揭示多样性。

在未来，我们将应用数据挖掘技术来分析纵向研究，这可能会揭示小群体的规模和特征随时间的变化。我们还将从挖掘的角度分析现有的调查设计方法，这可能会导致一些建议，以收集更鲁棒和有意义的数据集用于聚类。

致谢

感谢QTEMA为我们提供调查数据。特别感谢布莱金理工学院的Robert Feldt教授和Tony Gorschek教授分享数据并提供反馈。本研究由欧盟结构基金通过比较商业创新中心III (CBIC III)项目和the Knowledge资助

[11]Like “No Reply” (“Inget svar” in Swedish) was used by QTEMA which itself can be a category, for numerical values Null is the standard choice.

基础(KKS)通过项目20130085:测试关键系统特性(TOCSYC)。

9. 引用

[1] T. Xie, J. Pei, and A. Hassan, "Mining software engineering data," in Software Engineering - Companion, 2007. ICSE 2007 Companion. 29th International Conference on, May 2007, pp. 172–173.

[2] S. L. Pfleeger and B. A. Kitchenham, "Principles of survey research: Part 1: Turning lemons into lemonade," SIGSOFT Softw. Eng. Notes, vol. 26, no. 6, pp. 16–18, Nov. 2001. [Online]. Available: http://doi.acm.org/10.1145/505532.505535

[3] B. Kitchenham and S. L. Pfleeger, "Principles of survey research part 6: Data analysis," SIGSOFT Softw. Eng. Notes, vol. 28, no. 2, pp. 24–27, Mar. 2003. [Online]. Available: http://doi.acm.org/10.1145/638750.638758

[4] ——, "Principles of survey research: Part 5: Populations and samples," SIGSOFT Softw. Eng. Notes, vol. 27, no. 5, pp. 17–20, Sep. 2002. [Online]. Available: http://doi.acm.org/10.1145/571681.571686

[5] C. X. Ling and C. Li, "Data mining for direct marketing: Problems and solutions." in KDD, vol. 98, 1998, pp. 73–79.

[6] M. Gamon, A. Aue, S. Corston-Oliver, and E. Ringger, "Pulse: Mining customer opinions from free text," in Advances in Intelligent Data Analysis VI. Springer, 2005, pp. 121–132.

[7] S. Morinaga, K. Yamanishi, K. Tateishi, and T. Fukushima, "Mining product reputations on the web," in Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2002, pp. 341–349.

[8] B. Liu and L. Zhang, "A survey of opinion mining and sentiment analysis," in Mining text data. Springer, 2012, pp. 415–463.

[9] B. A. Kitchenham, S. L. Pfleeger, L. M. Pickard, P. W. Jones, D. C. Hoaglin, K. El Emam, and J. Rosenberg, "Preliminary guidelines for empirical research in software engineering," Software Engineering, IEEE Transactions on, vol. 28, no. 8, pp. 721–734, 2002.

[10] J. Moses, "Benchmarking quality measurement," Software Quality Journal, vol. 15, no. 4, pp. 449–462, 2007.

[11] J. Moses and M. Farrow, "Tests for consistent measurement of external subjective software quality attributes," Empirical Software Engineering, vol. 13, no. 3, pp. 261–287, 2008.

[12] J. Moses, "Should we try to measure software quality attributes directly?" Software Quality Journal, vol. 17, no. 2, pp. 203–213, 2009.

[13] T. Gorschek, E. Tempero, and L. Angelis, "On the use of software design models in software development practice: An empirical investigation," Journal of Systems and Software, vol. 95, pp. 176–193, 2014.

[14] D. J. Hand, "Data mining: statistics and more?" The American Statistician, vol. 52, no. 2, pp. 112–118, 1998.

[15] M. J. Shaw, C. Subramaniam, G. W. Tan, and M. E. Welge, "Knowledge management and data mining for marketing," Decision support systems, vol. 31, no. 1, pp. 127–137, 2001.

[16] R. Mikut and M. Reischl, "Data mining tools," Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, vol. 1, no. 5, pp. 431–443, 2011. [Online]. Available: http://dx.doi.org/10.1002/widm.24

[17] D. Fasulo, "An analysis of recent work on clustering algorithms," Department of Computer Science & Engineering, University of Washington, 1999.

[18] T. Moon, "The expectation-maximization algorithm," Signal Processing Magazine, IEEE, vol. 13, no. 6, pp. 47–60, Nov 1996.

[19] C. B. Do and S. Batzoglou, "What is the expectation maximization algorithm?" Nature biotechnology, vol. 26, no. 8, pp. 897–900, 2008.

[20] C. Ordonez and P. Cereghini, "Sqlem: Fast clustering in sql using the em algorithm," in Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data, ser. SIGMOD '00. New York, NY, USA: ACM, 2000, pp. 559–570. [Online]. Available: http://doi.acm.org/10.1145/342009.335468

[21] J. A. Hartigan and M. A. Wong, "Algorithm as 136: A k-means clustering algorithm," Applied statistics, pp. 100–108, 1979.

[22] M. Hall and G. Holmes, "Benchmarking attribute selection techniques for discrete class data mining," Knowledge and Data Engineering, IEEE Transactions on, vol. 15, no. 6, pp. 1437–1447, Nov 2003.

[23] M. Hassan, "□aAIJintelligent□ yardstick□ aA□I,・an approach of ranking to filter non-promising attributes from schema in data mining process," in Intelligent Control and Automation, ser. Lecture Notes in Control and Information Sciences, D.-S. Huang, K. Li, and G. Irwin, Eds. Springer Berlin Heidelberg, 2006, vol. 344, pp. 623–632. [Online]. Available: http://dx.doi.org/10.1007/978-3-540-37256-1 79

[24] R. Kohavi and G. H. John, "Wrappers for feature subset selection," Artificial intelligence, vol. 97, no. 1, pp. 273–324, 1997.

[25] M. A. Hall, "Correlation-based feature selection for machine learning," Ph. D. dissertation, The University of Waikato, 1999.

[26] C. Fraley and A. E. Raftery, "How many clusters? which clustering method? answers via model-based cluster analysis," The computer journal, vol. 41, no. 8, pp. 578–588, 1998.

[27] Weka. (2014) Em. [Online]. Available: http://weka.sourceforge.net/doc.dev/weka/clusterers/EM.html

[28] A. P. Dempster, N. M. Laird, D. B. Rubin et al., "Maximum likelihood from incomplete data via the em algorithm," Journal of the Royal statistical Society, vol. 39, no. 1, pp. 1–38, 1977.

[29] J. N. Breckenridge, "Validating cluster analysis: consistent replication and symmetry," Multivariate Behavioral Research, vol. 35, no. 2, pp. 261–285, 2000.

[30] QTEMA. (2014) Qtema official homepage. [Online]. Available: http://www.qtema.se

[31] M. R. Anderberg, "Cluster analysis for applications," DTIC Document, Tech. Rep., 1973.