Applying clustering to analyze opinion diversity

Mohammad Mahdi Hassan
Computer Science Department
Karlstad University
Karlstad, Sweden
mohammad.hassan@kau.se

Martin Blom
Computer Science Department
Karlstad University
Karlstad, Sweden
Martin.Blom@kau.se

ABSTRACT

In empirical software engineering research there is an increased use of questionnaires and surveys to collect information from practitioners. Typically, such data is then analyzed based on overall, descriptive statistics. Even though this can capture the general trends there is a risk that the opinions of different (minority) sub-groups are lost. Here we propose the use of clustering to segment the respondents so that a more detailed analysis can be achieved. Our findings suggest that it can give a better insight about the survey population and the participants' opinions. This partitioning approach can show more precisely the extent of opinion differences between different groups. This approach also gives an opportunity for the minorities to be heard. Through the process significant new findings may also be obtained. In our example study regarding the state of testing and requirement activities in industry, we found several significant groups that showed significant opinion differences from the overall conclusion.

Keywords

Empirical Survey, Clustering, Data Mining, Partitioning, Grouping, Diversity, Minority, Expert Opinion.

1. INTRODUCTION

In the process of software development many different forms of data is produced. Traditional forms of data are as follows [1]- 1) Code bases, 2) Execution traces, 3) Historical code changes, 4) Bug databases etc.

In the recent past, huge investments have been poured into software process automation as it can reduce development cost and increase product quality. Process automation not only produces some traditional forms of data in large amounts, but also gives a chance to store and extract new forms of data. Some of the other forms of software engineering data can be described as follows:

- Test cases - Heavily used in automated testing processes. Test cases can be generated manually or through an automated process.
- System build traces - Component building and their integration process has become highly automated and can hence be traced more easily.
- Team and personal data - Commercial tools exist to collect and trace developers' (as well as teams') effort and working patterns.
- Development process data - Tools also exist to collect complete development process data.

In this paper, we analyze a traditional data source, "Opinion Survey", which is generally not considered for data mining (DM). Our study suggests this form of data may have some potential patterns which can be extracted through a clustering process. It may reveal new information and attract attention to alternative perspectives.

Collecting survey data from software development practitioners to analyze statistically is one of the key areas of software engineering research [2]. In recent time online facilities and tools make it easier to collect survey opinions in a frequent manner, so a considerable amount of survey data is present in most of the software organizations.

We notice that most of the survey analyses are performed using traditional statistical methods and measures (like mean, median, variance and some data analysis tests) for their findings [3]. Overall they consider the whole survey population as a single group with some sampling techniques to extract varieties [4]. In some cases the population is also partitioned into sub groups based on some background information [3]. That does not reveal opinion diversity properly as similar opinions can exist in different segments of the population, whereas people within the same group might have different opinions.

In this study, we apply clustering techniques on opinion survey data which were collected in a categorical or numerical form. Clustering without any perceived bias divides the population into different clusters of sub- populations which to some degree have a similar opinion. There are some benefits and opportunities we observe using this approach such as:

- It can reduce manipulation in grouping, as it generates groups based on their opinion. Moreover background information can also be incorporated with opinions, if necessary, when clustering is applied.

EASE '15, April 27 - 29, 2015, Nanjing, China
Copyright 2015 ACM 978-1-4503-3350-4/15/04 ...$15.00.
http://dx.doi.org/10.1145/2745802.2745809

- It can exhibit opinion difference in the population more precisely. Statistical variance [3] can only show overall agreement or disagreement, whereas grouping by DM (clustering) can show variance in each group and intragroup agreement and disagreement.

- It can identify minority groups which would not be identified otherwise. In most cases, minority groups lose their voices as results are presented in a more aggregated manner.

- Opinion difference and background information may reveal groups with distinct characteristics which may lead to generating valid hypotheses. In consequence further research can be designed to investigate those groups and related hypothesis.

- In some cases, certain forms of correlation between different aspects of opinion are only visible within a cluster and are not obvious until cluster formation.

To investigate the clustering approach we used a survey conducted by a Swedish consultant company. Based on the survey they had come to some overall conclusions like status of testing is satisfactory but not the requirement process. After applying clustering in our study survey we found some sizeable groups within the survey population whose opinions are different than the overall conclusion both in intensity as well as in direction.

We organize our paper as follows: Section II contains related work, in Section III we try to answer some questions that might arise in the clustering process, in Section IV we describe the background of the sample survey which we have used in our study, in Section V we describe the process of applying clustering on categorical data, Section VI is dedicated to numerical data clustering, in Section VII we further discuss our approach. Finally, we conclude in Section VIII.

2. RELATED WORK

According to our investigation on “expert opinion survey data analysis using data mining techniques”, there is a lack of research in software engineering in this regard. We have found that mining research has been done in marketing [5] and business related fields on survey data [6]. Most of those data are from ordinary people; due to the internet boom it has become easier for the business organizations to collect customersâĂŹ opinions through web forms [7] as well as traditional methods (phone, paper based etc.) and preserve them in digital form. The key goal for them is to identify potential customers and products, which ultimately increases their business and revenues [8].

Opinion survey on practitioners is one of the core research practices in the field of software engineering. There are some standard guidelines to analyze such survey data, which basically are some rational investigation methods based on a simple statistical approach. Barbara A. Kitchenham [3] & [9] describes some of those methods with a caution for using advanced statistical methods like Bayesian analysis. They mentioned that âĂIJBayesian methods are not usually used in software engineering studiesâĂİ and recommends to get help from statisticians. In recent times John Moses [10], [11] & [12] has proposed a software quality prediction model based on expert opinion using Bayesian inference and Markov Chain Monte Carlo (MCMC) simulation. In general descriptive statistical techniques as well as hypothesis tests are used to analyze survey opinion [13].

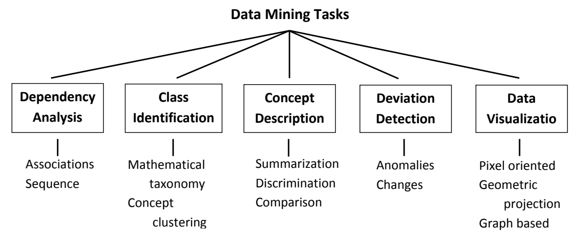

Figure 1: A taxonomy of data mining tasks[15]

There is a kind of uneasiness we can observe in the survey research domain towards using advanced statistical models. The situation is also true for machine learning techniques, which are based on advanced mathematical and statistical models [14]. Due to the small amount of data from those surveys, researchers are also discouraged to use DM techniques for analysis. In subsequent sections, we show that applying DM on survey data has a good chance to discover different perspectives which may be overlooked and not discovered using traditional rational and simpler statistical analyses.

Prior to describing our main study, we give a brief overview of related data mining tasks used in our analysis process. According to Shaw et. al. [15], DM tasks can be divided into following components (See also Figure 1) - 1) Dependency Analysis, 2) Class Identification, 3) Concept Description, 4) Deviation Detection, 5) Data Visualization.

Basically, we performed group identifications (Class Identification) and characterization (Concept Description) on sample survey data. Details of the analysis process will be described in later sections.

3. CLUSTERING - SOME RELATED QUESTIONS

We use data mining techniques on survey data for quantitative analysis, then we use the result produced by clustering to find new facts or interpret some phenomena which suggest opinion diversity. The following questions guided us in the process:

- Which tool should we use?

- Which clustering algorithms should we use?

- How many questions should we use for mining?

- How many clusters?

- How to analyze clusters individually?

In the following sections we will try to answer those questions.

3.1 Which tool should we use?

There are many commercial as well as open source data mining tools available [16]. Some mining tools provide rich presentational kits that can be useful to comprehend the clustering result. We used Weka (`http://www.cs.waikato.ac.nz/ml/weka/`) an open source data mining tool. It contains prominent clustering algorithms and provides a simple visualization interface. It also supports exporting results after clustering which can be imported (with minor processing) to other statistical tools for further analysis.

3.2 Which Clustering algorithms should we use?

There are basically two kinds of clustering algorithms, Partitioning (or K-clustering) and Hierarchical clustering [17]. Depending on the target of analysis one should first choose a type of clustering paradigm, then choose a specific algorithm for further processing. Selecting a suitable algorithm may have an impact on findings and their interpretation. There are several factors one should consider while selecting an algorithm; data type (numeric, categorical etc.), size of the dataset and characteristics of the dataset (like rate of missing data, presence of outliers and distribution of data points etc.).

We used the Expectation Maximization (EM) [18, 19] algorithm for clustering. It is a method of cluster analysis which aims to classify n observations (data points - in our case opinion records) into k clusters (groups) in which each observation belongs to a particular cluster based on their likelihood. EM is a robust algorithm as it can handle missing data which is common in opinion surveys. EM algorithm is based on strong statistical basis which is highly scalable and linear in database size. It can handle high dimensionality and expected number of clusters can be set manually. With a good initialization it converges fast [20]. Besides that, both categorical and numerical data can be mined using EM, something which is not supported by some other popular clustering algorithms like K-means algorithm [21].

3.3 How many questions should we use for mining?

In general each survey question can be considered as a possible attribute for clustering. Attribute selection is a challenging problem in data mining [22, 23]. We can reduce the computation cost by considering relevant attributes only. It also helps to extract information in a better shape [22]. There is no specific limit for number of clustering attributes; rather it depends on nature of the dataset. Some attribute selection techniques like Wrapper Subset Evaluation [24], CFS (Correlation-based Feature Selection) [25] and PCA (Principal component analysis) [22] might help in this regard. Identifying the research interest may also guide in this regard. Beside that, due to the small sample size, survey opinion data is vulnerable to producing inconclusive clusters in case of using too many questions in clustering. In our case we only use two to three questions in the clustering process to partition the dataset.

3.4 How many clusters?

Llacking prior knowledge of the dataset it is difficult to predict the possible number of clusters [26]. In clustering algorithms like EM we can set a prior number so that dataset will be partitioned accordingly. There are some statistical measures like simple distributions for the cluster models [27][1], standard deviation (for numerical data), log and maximum likelihood [18, 28] etc. that can help us to determine the possible number of clusters. We will discuss it in more detail with an example in the following sections.

3.5 How to analyze clusters individually?

Choosing suitable indicators to present the cluster's characteristics is important. In most cases standard statistical indicators like percentage, count, mean/median variance etc., are used for this purpose. The choice of indicators can influence while observing the differences between clusters. We used clustering to partition the dataset so that similar opinions are regrouped into clusters. We expect that those smaller groups will be more cohesive and will show some opinion variations. Initially we investigate each cluster separately and compare them with each other as well as with the general population using statistical indicators for each attribute (i.e. each survey question). Through this process we can also observe some correlation between different attributes. A significant correlation between dependent attributes (i.e. used in clustering) vs independent attributes (i.e. not used in clustering) might indicate the reason of opinion difference.

3.6 Can we repeat the clustering process?

Clustering process is highly repeatable if the parameter settings as well as the dataset are intact. In our case we use the default parameter settings provided by Weka including the initial random seed, we only changed the expected number of clusters value. To support replication of the study, one can provide value of the parameter settings[2]. Replication across different datasets is possible but challenging as clusters in different datasets might be inconsistent [29].

4. SAMPLE SURVEY OVERVIEW

Data was collected from a number of Swedish IT-companies through an online questionnaire that was distributed by an IT education and consulting company, QETMA [30], as part of their work to assess the current state-of-practice in the IT-industry. The questionnaire included 21 questions ranging from background questions on personal and corporate data to questions on various technical and non-technical aspects related to IT and the processes used by the participants.

The questionnaire was distributed annually and this study is based on the questionnaire from 2010. It was answered by 153 respondents. We chose to include both categorical and numerical data to evaluate the data mining methodology for both types of questions. The questions used for clustering are listed below[3], together with the answers and the percentage of respondents who selected that answer[4]:

1. “Which of the following sentences best describes what development methodology you use most often?”

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

[2] Following parameters is used in WekaâĂŹs EM implementation âĂŞ 1) Maximum number of iterations = 100, 2) Minimum 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 patterns. For larger datasets there might be a need for a systematic attribute reduction process to select clustering attributes.

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

(a) A traditional, plan driven process with requirements followed by (in order) design, implementation and testing, 30%
(b) An agile process where both requirements, implementation and testing work is done in short iterations, 14%
(c) A conscious mix of a traditional, plan driven process and an agile process using short iterations, 52%
(d) Others, 4%

2. “Does your company / unit have a functioning organization and process for working with requirements?”
   (a) To a very low degree, 10%
   (b) To a low degree, 45%
   (c) To a high degree, 35%
   (d) To a very high degree, 8%

3. “Does your company / unit have a functioning organization and process for working with testing / verification / validation?”
   (a) To a very low degree, 3%
   (b) To a low degree, 26%
   (c) To a high degree, 56%
   (d) To a very high degree, 15%

4. “How much time (in %) do you spend on requirements?”[5]
   (a) <20%, 60%
   (b) >30%, 3%
   (c) mean, 15%

5. “How much time (in %) do you spend on test, verification and validation?”
   (a) <20%, 31%
   (b) >30%, 12%
   (c) mean, 22%

The selection of these particular questions out of the 21 possible questions was based on the profile of QTEMA that focuses on requirements and testing and the possible correlation between those fields and the development method used. Questions 1 to 3 were used for categorical data clustering and the remaining two were used for clustering on numerical data. For simplicity we will use Traditional, Agile and Blend as a replacement answers of question 1 for options a), b) and c) respectively.

We also included some other questions from the study that we deemed interesting and were related to the working habits of the respondents. These questions and the answer alternatives used in the survey are listed below together with their related statistics.

6. “How much experience do you have working with your current tasks?”
   (a) <1 years, 6%
   (b) 1-3 years, 18%
   (c) 3+ years, 76%
   (d) 5+ years, 56%
   (e) 10+ years, 33%

7. “Do you have the required competence to work professionally?”
   (a) To a very low degree, 0%
   (b) To a low degree, 4%
   (c) To a high degree, 56%
   (d) To a very high degree, 38%

8. “Which of the following best describes when in a typical development project you meet the customer?’
   (a) Never, 7%
   (b) In the beginning, 5%
   (c) At the end, 4%
   (d) At the beginning and the end, 21%
   (e) Continuous and on several occasions throughout the project, 54%
   (f) Daily throughout the project, 8%

9. “How much time (in %) do you spend on implementation / coding?”
   (a) <20%, 79%
   (b) mean, 34%

10. “How much time (in %) do you spend on design / analysis / modelling?”
   (a) <20%, 16%
   (b) mean, 15%

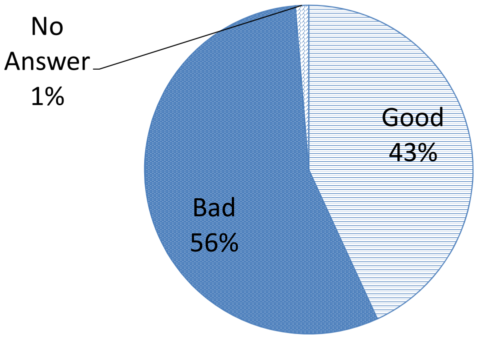

Figure 2: Overall requirement status

Based on the survey Qtema suggested that the condition of requirement related actives are in bad shape but overall test activates are satisfactory[6]. We present their overall conclusion in Figure 2 and Figure 3 respectively.

[5] This and the following questions have been separated from a compound question on time consumption and been rewritten to increase readability. The original question was -“How much time (in %) do you spend on the following activities? (For each activity you can state a number between 0-100, but the sum for all activities should be 100)”. Since the respondents could enter numbers freely in these questionnaires, we made categories(based on cut points). Beside presenting percentage of respondents , we also provide the mean value.

[6] Here “To a very low degree” and “To a low degree” are considered as bad whereas “To a very high degree” and “To a high degree” are considered as good

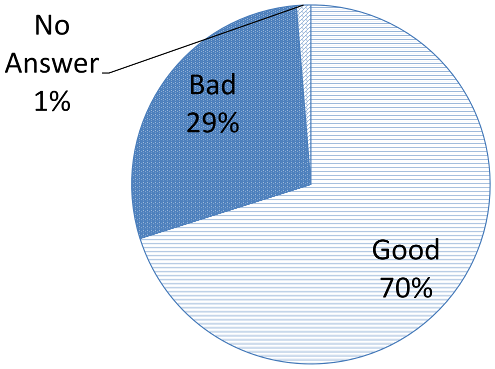

Figure 3: Overall test status

5. CLUSTERING CATEGORICAL DATA

Table 1: Pivot Table for Survey Question 1 to 3. "TVL": To a very low degree, "TL": To a low degree, "TH": To a high degree, "TVH": To a very high degree,

	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

There is an inherent support for clustering categorical data, and as the options are limited one can predict the maximum possible groups based on a combination of all categories. If the number of columns (each column contains opinions of a single question) used in grouping is N and each column respectively has $m_1, m_2 \ldots m_n$ number of categories then the maximum number of possible groups can simply be calculated as in equation 1 below:

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

In Table 1 we show the pivot table based on survey questions 1 to 3 which represent absolute categorization. Here out of forty eight possible groups only five have a population of more than or equal to ten, which suggests the whole population becomes fragmented into insignificant groups. So in reality we may find that a great number of those possible groups due to absolute categorization are empty or insignificant, and that only a few significant groups exist.

Naturally for categorical data the number of possible groups increases with an increase in the number of questions. In some surveys numerical values are also used to express opinions (like in scale questions) which increases the combination of possible options dramatically. So in most cases it is infeasible to observe and analyze all possible groups individually. Besides that in a more complex scenario we might find that in some segments of the population some categories are obsolete but for others they are not. In such scenario absolute categorization might mislead the analysis process. On the other hand our interest is not all groups but some groups who have different opinions. There exist many clustering algorithms which can handle situations like this more easily by identifying significant groups.

In following sections we will describe the process of applying clustering on our sample survey.

Table 2: Cluster distribution, number of cluster set as 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

5.1 Initial step

We used survey questions (SQ) one to three for categorical clustering (see section 4). Initially we set expected cluster as one and in Table 2 we show the clustering results provided by Weka's EM implementation[7]. The result obviously suggests that for each question there are significant opinion diversities (like in SQ2 for each possible answer there are more than ten people). In the following steps we will try to partition the overall population into different segments so that in each segment diversity is reduced and similar kinds of opinions (in all three questions) reveal themselves distinctively.

5.2 Partitioning

In Weka's EM implementation partitioning can be done in two ways: 1) With default settings where the expected number of clusters is set to -1 which suggests EM will automatically determine the number[8], 2) Set an expected number of clusters and observe the results.

In Table 3 we show the result of the default setting which produce four clusters. Here we use two indices to denote a cluster, like in C(-1,1) first index suggest the expected

[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]

cluster setting which is default -1, and second index indicate a specific cluster. We follow the same format across Table 3 to Table 8 to represent clustering results.

Table 3: Clusters distribution, number of cluster set as default -1

SQ	**Answers**	**C**	**C**	**C**	**C**
SQ	**Answers**	**(-1,1)**	**(-1,2)**	**(-1,3)**	**(-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

In the next step we start to investigate the clustering process using manual settings of expected number of clusters(for convenience we denote it as N). At first we set N as two. In Table 4 we present the results. Cluster C(2,2) suggests a new opinion as most of the people in the cluster are confident regarding requirement process (SQ2). On the other hand cluster C(2,1) does not look cohesive as there are multiple dominant choices for the same question within same cluster. So we continue to increase the value of N and analyze the results.

5.3 Identifying interesting clusters

Our goal is to observe alternative opinions among the population segments. So after getting the clustering results for each setting we try to locate those clusters which are significant in size and show some form of alternative conclusions. Across different clusters opinion difference is represented by population distribution among opinion categories in each question. We check each question (used in clustering) separately for each of the clusters to observe the difference. Like in table 3 out of four clusters two are not significant in size, C(-1,2) and C(-1,4) are 6% (9 persons) and 2% (3 persons) of the population respectively. On the other hand size of the C(-1,1) and C(-1,3) are 28%(43 persons) and 64%(98 persons) of the population respectively. Here cluster C(-1,3) align with the general conclusion(low confidence on requirement process) whereas cluster C(-1,1) has a different opinion compared to the general population as an overwhelming majority responded positively in SQ2 (higher confidence on requirement process).

Table 4: Clusters distribution, number of cluster set as 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

Table 5: Clusters distribution, Interesting Clusters through the whole mining process

SQ	**Answers**	**C**	**C**	**C**	**C**	**C**
SQ	**Answers**	**(5,1)**	**(6,6)**	**(6,2)**	**(6,4)**	**(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**

After locating interesting clusters we separate them from rest of the population and analyze them individually. The same cluster can appear in different settings, with an increase of N either those clusters reappear or disintegrate into smaller clusters. If we find a set of clusters which emerge in different settings of N but exhibit common features in all the attributes used for clustering, then we consider them as different representations of the same group. In table 5 we show related distribution stats of such groups. Here we put the results based on the settings of N where the representative cluster of each group is largest in size.

5.4 Stop clustering

As our interest is to locate alternative opinions, finding a group is more important than overall categorization. We continued to increase the value of N and looked for interesting clusters even though the overall categorization deteriorates (see the log likelihood for different settings of expected clusters in Figure 4). After setting the value of N as seven, we did not find any significant clusters that present a new opinion. In Figure 5 we show the change of size of the different groups in different settings of N.

5.5 Analyze significant groups

We used all survey questions mentioned in section 4 to analyze each of the groups we extracted through the clustering process[9]. As expected all of them show some distinction in SQ1 to SQ3 as they were used in clustering, but they also

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

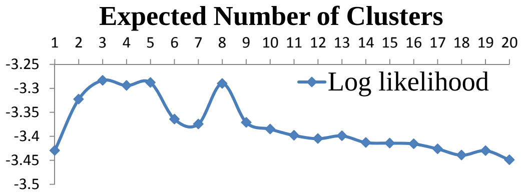

Figure 4: Log likelihood for different settings of expected number of clusters

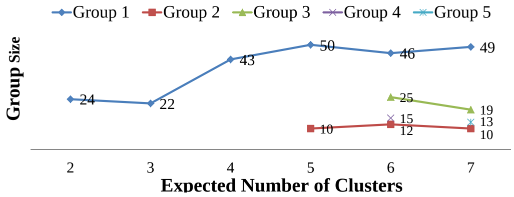

Figure 5: Change of Size of different groups showed in Table 5

show some interesting differences in the rest of the questions. In the following section we will present each of the groups and describe some of their distinctive features.

1. Group 1: This is the largest group with an alternative opinion where the majority of the population are confident regarding requirement as well as testing process. Some distinctive properties of this group:

   (a) 94% show confidence in the requirement process and 98% show confidence in the testing process (i.e. chose either *c* or *d* in SQ2 or SQ3). Note, in the general population 43% and 70% showed confidence in the requirement and the testing process respectively.

   (b) 76% suggest some form of frequent interaction with the customer (chose *e* or *f* in SQ7) whereas in the general population it is 62%.

2. Group 2: This is a small group where all members are highly confident regarding requirements as well as the testing process. Some distinctive properties of this group:

   (a) Except one member who did not answer, all chose *d* in SQ2 or SQ3.

   (b) They are highly experienced; seven out of twelve members (58%) have more than ten years experience, in the general population it is 33% (SQ7).

   (c) They are highly confident regarding their professional competence; eight out of twelve members (67%) chose *d* in SQ8, in the general population it is 33%.

3. Group 3: This group not only feels less confident about the requirement process but also feels the same regarding the testing process. Some distinctive properties of this group:

   (a) 100% of the members chose *b* in SQ3 (suggests less confidence in testing process) which is only 29% (combined *a* and *b*) in the general population.

   (b) In SQ5 they have a mean (16%) which is much lower than the general population (22%). It might suggest that they are less involved in the testing process.

4. Group 4: This is similar to group 3 but smaller in size and populated by people who follow a traditional development process. Some distinctive properties of this group:

   (a) 100% of the members chose *a* in SQ1.

   (b) 100% of the members show less confidence regarding the testing process.

   (c) 60% of the members suggest they do not frequently interact with the customer which is much higher compared to the general population (37% chose *a* to *d* in SQ8).

5. Group 5: This group aligns with group 1, but is smaller in size and populated by people who follow a traditional development process. Some distinctive properties of this group:

   (a) 100% of the members chose *a* in SQ1.

   (b) 100% of the members show confidence both on requirement and the testing process.

   (c) For SQ5 they have a higher mean compared to the general population (29% vs 22%) which might suggest they are more involved in testing process.

It might be possible to find more distinctive properties by considering other survey questions which we have omitted. On the other hand in this paper our main purpose is to show that strong alternative opinions exist and that clustering techniques can reveal them.

6. CLUSTERING NUMERICAL DATA

In a survey numerical data is used to capture information in many different forms. Sometimes it is used to collect direct information like the size or value, other times it is used to collect relative information through some form of numerical scale, like in SQ4 and SQ5 where information was collected as a percentage. In both forms of the numerical data we can apply clustering techniques to identify some cohesive groups which might show some distinct characteristics.

In the following sections we will describe the process of numerical clustering.

6.1 Initial Steps

We used SQ4 and SQ5 mentioned in section 4 to experiment with numerical data. Both of the questions are related to the working habits of the participants.

First we set the value of *N* as 1 to get related stats for the general population. We show the results in Table 6[10] in the 3rd column. Both of the questions have a very high standard deviation which suggests there might be significant differences in working habits among the survey participants.

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

Table 6: Clustering Result of Numerical Data with default settings

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**

Through clustering we will try to reveal groups that are different from each other according to their working habits (in this case testing and requirements only). Later we will analyze them individually to observe possible opinion differences due to the working habits.

6.2 Partitioning

Like for categorical data Weka also supports two modes of clustering numerical data: 1) A priori setting of expected number of clusters N, 2) Default value of N as -1 to get the number of clusters through cross validation. Presentation of results after the clustering process is different compared to categorical clustering, as it presents the results in the form of mean and standard deviation for each cluster.

We start with default settings which produce two clusters C(-1,1) and C(-1,2) shown in Table 6. Both clusters showed higher standard deviations; as we are looking for cohesive groups we continue further clustering.

6.3 Identifying interesting clusters

We identify interesting clusters based on mean and standard deviation. We expect that a cohesive group will have much lower standard deviation in each of the attributes used in clustering (compared to the general population). We also expect that each of those groups will have different means in some attributes which will make them distinct. After identifying the group, we observe and analyze the group across different settings of N to select the cluster which best represents the group. Here we mainly consider standard deviation and size, so that the group is large enough with an acceptable standard deviation.

In Table 7 we show a group (we denoted it as Group 1) which has a lower but consistent mean in both SQ4 and SQ5 compared to the general population for different settings of N. Here C(11,7) is the most cohesive considering the standard deviation, whereas C(8,1) is the largest with an acceptable standard deviation. Here we refrain from giving any hard rule to choose any specific cluster, rather our purpose is to show that we can use clustering to identify such groups. In the analysis phase we will show that both of the clusters represent a very similar group considering their opinions in other survey questions.

6.4 Stop clustering

In Figure 6 we show the comparison of the number of clusters with a size of 10% of the population for different value of N. It is obvious from the figure that the number of clusters with a significant population is stable and goes up to five initially then decreases to three when the value of N is high. On the other hand initially there were no smaller groups but after their first emergence, their number is almost linear to N.

Table 7: Representative clusters of Group 1 across different settings of N

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**

As we increase the value of N we start to get smaller clusters, with a high value of N the chance of getting a new distinct group with significant population is small. Hence, we can stop the clustering process.

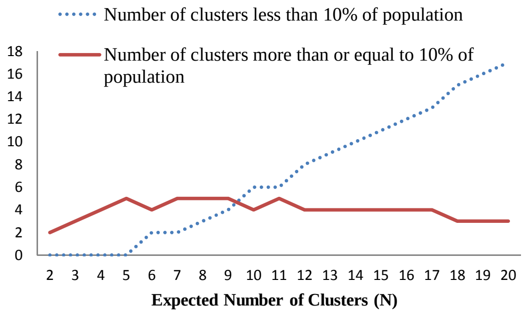

Figure 6: Expected Number of Clusters (N) vs Number of Clusters which contain more or less than 10% populations

6.5 Analyze significant groups

We already mentioned that a group can be represented by different clusters for different settings of N. In Table 7 we show different clusters that represent group 1. First we will analyze C(8,1) and C(11,7) the largest and the most cohesive cluster respectively representing group 1. Later we will analyze two more interesting groups shown in Table 8.

1. Group 1: This group represents people who are significantly less involved in requirement activities (in SQ4 the mean value is around 9 which is 40% less compared to the general population). They also have a slightly lower involvement in testing activities.

   (a) C(8,1): This is the largest cluster representing group 1. In this cluster 66% of the members showed less confidence in the requirement process which is higher than the general population (56%). They also less frequently interact with the customer (52% vs 62% in SQ8).

   (b) C(11,7): This is the most cohesive (considering standard deviation) cluster representing group 1. In this cluster 75% member showed less confidence in the requirement process which is much higher than the general population. They also less frequently interact with the customer (50% vs 62% in SQ8).

Table 8: Group 2 and 3

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**

So besides having similar means in SQ4 and SQ5, both of the clusters share a similar tendency in opinions which we had anticipated.

2. Group 2: This group represents the segment of the survey participants who are more involved in testing compared to the general population. They showed a significantly higher confidence regarding the testing process compared to the general population (88% vs 70% in SQ3). They also showed higher confidence on requirement process (54% vs 43%).

3. Group 3: This group represents survey participants who are more involved in requirement related work (in SQ4 they have a mean of 28 which is 87% higher than the general population). They are more satisfied with the requirement process compared to the general population (63% vs 43% in SQ2). They also have much higher interaction with customers (94% vs 62% in SQ8).

7. DISCUSSION

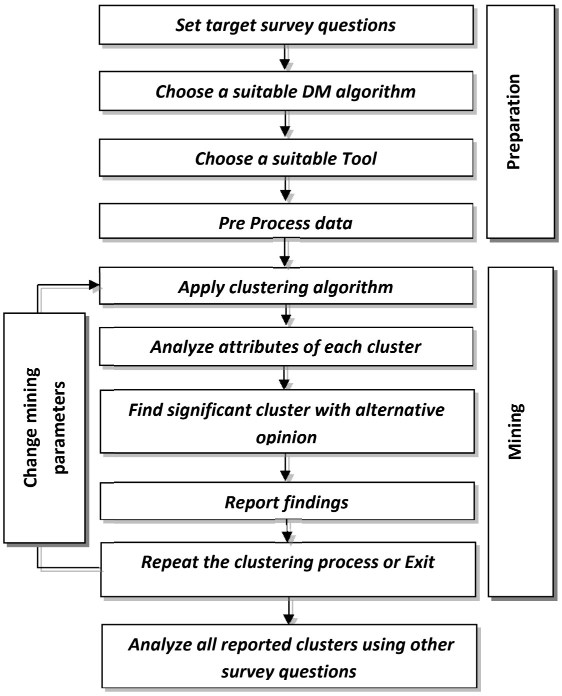
Set target survey questions
Choose a suitable DM algorithm
Choose a suitable Tool
Pre Process data
Preparation
Apply clustering algorithm
Analyze attributes of each cluster
Find significant cluster with alternative opinion
Report findings
Repeat the clustering process or Exit
Change mining parameters
Mining
Analyze all reported clusters using other survey questions

Figure 7: Process Diagram

Anderberg showed that it is extremely difficult to comprehend possible groupings from a dataset by simple human ability. He gave an example where a possible grouping of 25 observations into 5 groups is huge (exactly 2,436,684,974,110,751) [31]. So even for a small survey it is almost impossible to exhaustively partition the population manually and investigate their characteristics. But clustering can help in this regard as it is used in other domains to solve similar problems.

In this study we have applied clustering techniques in a systematic way to partition the survey population, then separate and analyze significant groups who showed alternative opinions. In Figure 7 we show the process using a flow graph. We already discussed the process with examples in previous sections. Here we discuss some important factors which might impede the process.

Preparing data is an important activity before starting the mining process. Some algorithms are not robust enough to handle missing data so empty records need to be weeded out or to be filled with some meaningful data to separate them from others[11]. In the current study survey we saw very few participants refrain from answering questions (Like only two participants did not answer SQ1).

In some cases participants induce nonstandard information based on their own understanding. Like in this study survey, 4 out of 153 participants provided process names in SQ1 which were not in the list (put under category "Other"). In this study we did not change their provided data, but in some other scenario if the number of such cases is high then revising them might lead to better clustering.

8. CONCLUSION AND FUTURE WORK

Application of data mining in opinion based surveys is not common in software engineering; probably the small number of survey participants may discourage researchers to use DM as an analysis tool. On the other hand within the survey population there might exist some smaller groups who may have a significant perspective difference that can lead to future success or warn against failure. In traditional methods it is difficult to observe opinion diversity among different groups as finding them is challenging. So in most of the analyses some form of overall statistical indicators are used to get a sort of comprehensive understanding, which undermine opinion diversity. In our study we show that some common clustering tools and techniques can help us to reveal diversity.

In the future we will apply data mining techniques to analyze longitudinal studies which might reveal the size and change of characteristics of minor groups over time. We will also analyze existing survey design approach from mining point of view, which may lead to some recommendations to collect more robust and meaningful dataset for clustering.

Acknowledgments

Thanks to QTEMA for providing us the survey data. Special thanks to Professor Robert Feldt and Professor Tony Gorschek of Blekinge Institute of Technology for sharing the data and providing feedback. This research was funded by EU Structural Funds through the Compare Business Innovation Centre III (CBIC III) project and by The Knowledge

[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.

Foundation (KKS) through the project 20130085: Testing of Critical System Characteristics (TOCSYC).

9. REFERENCES

[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, "âĂIJintelligent yardstickâĂİ, 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.