

How to Support Customer Segmentation with Useful Cluster Descriptions

Hans Friedrich Witschel^(✉), Simon Loo, and Kaspar Riesen

University of Applied Sciences Northwestern Switzerland (FHNW),
Riggenbachstrasse 16, 4600 Olten, Switzerland
{HansFriedrich.Witschel,Kaspar.Riesen}@fhnw.ch,
Simon.Loo@students.fhnw.ch

Abstract. Customer or market segmentation is an important instrument for the optimisation of marketing strategies and product portfolios. Clustering is a popular data mining technique used to support such segmentation – it groups customers into segments that share certain demographic or behavioural characteristics. In this research, we explore several automatic approaches which support an important task that starts *after* the actual clustering, namely capturing and labeling the “essence” of segments. We conducted an empirical study by implementing several of these approaches, applying them to a data set of customer representations and studying the way our study participants interacted with the resulting cluster representations. Major goal of the present paper is to find out which approaches exhibit the greatest ease of understanding on the one hand and which of them lead to the most correct interpretation of cluster essence on the other hand. Our results indicate that using a learned decision tree model as a cluster representation provides both good ease of understanding and correctness of drawn conclusions.

1 Introduction

In order to optimise their marketing strategies, companies need to understand the needs and preferences of their customers closely. Customer segmentation is a technique that allows companies to group customers into segments that share certain characteristics such as preferences or demand [16]. Based on customer segments and an understanding of their meaning, product offerings and marketing strategies can be better targeted by distinguishing certain categories of needs.

Many companies have begun to see the potential in gaining competitive advantage by extracting knowledge out of the abundant data that they can collect about their customers’ background, interests and behaviour. For instance, using data mining methods, they have been able to understand their customers’ purchasing habits better. The prevalent data mining method for customer segmentation is clustering (see e.g. [12]). Clustering is used to divide objects – e.g. customers – in the data set into clusters (group of related data points) such that objects within a cluster all share certain similarities. As a basis for clustering, each data object needs to be described by certain *attributes* or *features*.

The values of these attributes are compared between data objects to assess their similarity.

When segmenting customers of a company, an obvious and traditionally used family of attributes are demographic features, such as gender, nationality, family and socio-economic status. However, it has been recognised that attributes related to the interests and/or behaviour of customers can be more meaningful [13]. The resulting set of attributes usually has a mixture of types, including binary (e.g. interest in something, yes or no), categorical (e.g. nationality) and numeric attributes (e.g. number of times an event has occurred).

When clustering has been applied to a set of customers described by such attributes, the result is a set of clusters (or segments). For reasonably large customer bases, the size of such segments will be in the hundreds or thousands. Before a marketer can actually benefit from the result, (s)he needs to understand the “essence” of each segment, i.e. the characteristics that are shared by all customers within the segment and that make it different from the other segments. Usually, the marketer captures the essence of a segment by assigning a label, e.g. “rich singles who are active shoppers interested in accessories”.

Given the aforementioned size of segments, capturing and labeling the essence of clusters is not an easy task and needs to be supported by providing some sort of automatically generated cluster descriptions. Current data mining tools offer only limited functionality: usually, they allow to visualise the univariate distribution of attribute values within clusters. The drawback of such functionalities is threefold. First, they fail to highlight which attributes are important for a cluster, second therefore tend to be quite tedious to analyse if there are many attributes and third, they do not capture multivariate effects, i.e. meaningful combinations of attributes.

Some research proposed alternative ways of describing the essence of clusters, but there has been – to the best of our knowledge – no systematic evaluation of the quality of such alternatives. Therefore, the goal of our research is to assess the quality of a selection of cluster description techniques – specifically for the application to customer segmentation – in terms of ease of understanding and correctness of drawn conclusions. That is, our primary question is “which description format will an end user find most efficient and effective in interpreting clustered data?” Here, we mean by “efficient” the time required by the end user to understand the description. By “effective” we mean how correctly the end user will understand the description. If a description leads to accurate interpretation of the clustered objects, then we can say that the description is “effective”, if it leads to incorrect conclusions, it is ineffective. That is, we assume that on the one hand, marketers will prefer certain description methods because they are easy to grasp. On the other hand, a preferred method may be easy to grasp, but lead to wrong conclusions about the essence of a cluster, e.g. because of an oversimplification.

In order to carry out our evaluation, we proceeded as follows: we analysed the most promising proposed alternatives for segment descriptions from literature, as described in Sect. 2. We then derived a set of hypotheses regarding

how accurately they describe cluster essence (or may lead to wrong conclusions, respectively), see Sect. 3. Subsequently, we designed an experiment to verify these hypotheses by selecting a data set of customers, clustering it with a standard clustering algorithm and representing it with the chosen alternative description methods. The hypotheses were translated into a questionnaire that was, together with the cluster descriptions, given to a number of test persons. The precise setup of the experiment is described in Sect. 4. We then analysed and coded the responses of participants and mapped out the results as described in Sect. 5. Finally, we were able to draw conclusions, see Sect. 6.

2 Related Work

Clustering is a very active area of research and a great variety of clustering algorithms exists – see [7] for an exhaustive overview. The general approach of most clustering algorithms is to first establish a measure of similarity between data objects and then to try to group them such that objects in the same cluster are maximally similar and objects belonging to different clusters are maximally dissimilar. As mentioned above, researchers have also applied clustering to the problem of customer segmentation in various ways and settings, e.g. [12, 15].

The topic of how to describe and summarise the “essence” of clusters has received far less attention than the clustering of data itself. A notable exception is the area of text mining, where various methods for describing document clusters have been proposed (e.g. [3, 10, 14]). In document clustering, documents are usually represented by high-dimensional vectors where each term/word occurring in the whole document collection forms a dimension. For each dimension, a numerical attribute is created which represents the degree to which the corresponding term describes the content of the document. Hence, popular cluster description methods in text mining rely on the intuition that clusters should be represented by those terms that occur frequently within the cluster’s documents, but rarely otherwise. A cluster description is then a set of terms.

When clustering customers, however, the situation is usually different: as explained in Sect. 1 above, customers are usually represented by a mixture of – comparatively few – binary, categorical and numerical attributes. For nominal attributes the intuition from the text mining area does not work.

The approaches to summarising clusters of structured data objects with mixed attribute types can roughly be distinguished into two directions:

- Approaches that **summarise the distribution of attribute values** within the cluster. Many data mining tools offer visualisation of such distributions. Often, however, this can also happen by exploiting summaries that are built into existing clustering algorithms and that are simpler and faster to inspect by a human. For instance, the popular k-means algorithm [9] uses so-called *centroids* of clusters. A centroid is a vector of attribute values where each value summarises the distribution of values of a given attribute for all cluster members. For numerical attributes, the centroid contains the arithmetic mean of all values, for categorical attributes, it contains the mode.

- Other clustering approaches use more verbose summaries of distributions, especially for categorical attributes. For instance, the COBWEB algorithm [4] – an instance of so-called *conceptual clusterers* – represents clusters by a set of conditional probabilities, namely $P(A_i = V_{ij} | C_k)$ where A_i is a categorical attribute, V_{ij} represents one of the values of attribute A_i and C_k is a cluster. This essentially maps out all frequencies of the values of a categorical attribute within a cluster. A similar representation – using plain frequencies instead of probabilities – can be obtained for the expectation maximisation (EM) algorithm [2], a fuzzy clustering algorithm of which k-means is a special case.
- The other class of approaches relies on **learning a classifier with human-interpretable model** that is able to distinguish between the induced clusters. For instance, in [5, 6], it is proposed to learn a decision tree from the clustered data. This means that one first clusters the data and then trains a decision tree classifier to predict the cluster for unknown data objects, i.e. using the cluster number of each data object as class attribute. The resulting decision tree can then be inspected by a human. An important characteristic of a decision tree is the way in which it arranges attributes: given a decision tree for cluster C_k classifying objects into either “ C_k ” or “*not* C_k ”, the top-most attribute of that tree is the one that contributes most to reducing the uncertainty about whether an object belongs to C_k or not. This means that usually the top-most attribute is the one that most captures the “essence” of the cluster. If a tree becomes too large to be inspected easily by a human, it can be *pruned* (see e.g. [11]) such that only the most important attributes are visible.

Although there is no explicit proposal in the literature, other classifiers with human-interpretable models could be used in the same way. For instance – as can be seen e.g. by the comparative evaluation of classifiers in [8] – rule learners also yield models that humans can easily understand, e.g. RIPPER [1]. In that case, the model consists of a set of interdependent rules of the form *if* $A_i = V_{ij}$ *and* $A_l = V_{lm}$ *and* ... *then* C_k (see Sect. 4 for an example).

Although these possibilities have been proposed in literature and some of them are surely used in practice today, there has been no systematic and empirical evaluation of the suitability of these approaches to the problem of capturing and labeling the essence of clusters. We have chosen to contrast k-means centroids – as a member of the first category of approaches – with decision tree and rule representation – as representatives of the second category. We are aware that more sophisticated approaches exist for the first category – but we found it important to evaluate centroids because of their popularity, and chose not to evaluate the other ones because of their greater complexity and the limited number of participants that we could recruit.

3 Hypotheses

Taking into account the different characteristics of the representation approaches described in the previous section, it is natural to assume that these characteristics will have an impact on the correctness of conclusions that a human draws when

inspecting the representations. In the following, we will discuss our expectations regarding that impact for our chosen representations (centroid, decision tree and rules, see last section) and derive hypotheses, to be tested in the empirical part of our work.

3.1 Centroid Representation

Analysing the characteristics of the centroid representation leads to the following assumptions: First, we note that a centroid summarises *numerical attributes* by arithmetic mean of all values within the cluster. To a person with a background in statistics, it is clear that the values of the cluster elements are not necessarily all close to the mean. There can be various reasons for this – for instance, there could be an outlier that is pulling the mean or the variance of this particular attribute could be large. However, we assume that marketers may not be very educated in statistics and that even persons who do have a reasonable education in statistics, might be (mis-)led by the arithmetic mean to believe that the mean is representative of a majority of the attribute distribution’s values. For example, when analysing an attribute such as the age of persons within a cluster, they will believe that most people in the cluster have an age close to the arithmetic mean. We phrase this as the following hypothesis:

H1: Given a cluster representation that summarises numerical attributes using the arithmetic mean of the attribute values in the cluster, a human analyst will be led to believe that most of the attribute values are close to the mean and hence believe that the cluster can be characterised by the mean value w.r.t. the given attribute.

Second, we note that a centroid summarises *categorical attributes* by the mode, i.e. the most frequent of all values of the attribute within the cluster. Now, if there are two or more values of the attribute’s value distribution with almost equal frequency within the cluster, this will not be realised. As an example, consider analysing the gender of persons. If 51 % of persons in a cluster are female and 49 % are male, the mode of the gender attribute will be female. However, it is wrong to conclude that the cluster consists primarily of females (the number of females being not much higher than that of males and equal to our overall expectation). From this, we derive another hypothesis as follows:

H2: Given a cluster representation that summarises categorical attributes using the mode of the attribute values in the cluster, a human analyst will be led to believe that the vast majority of cluster members has that attribute value and hence believe that the cluster can be characterised by the mode w.r.t. the given attribute.

3.2 Rule Representation

When analysing rules, one needs to be aware of two important facts: First, each rule only captures the characteristics of *some* cluster members and one needs

to unite all statements made by the rules in order to capture a comprehensive picture of the cluster. Second, all statements about attribute values made within a rule need to be taken together, i.e. interpreted in a conjunctive way. Since this is a rather complex matter, we may assume that even if we instruct human analysts about it, there is a risk that they take statements from a single rule in isolation and generalise them to the whole cluster.

H3: Given a rule cluster representation and an explicit question concerning an attribute-value combination that is used within only one or two rules and that is not necessarily predominant in the whole cluster, a human analyst will state that in fact the combination is representative for all cluster members.

3.3 Decision Tree Representation

An important characteristic of a decision tree is that the attributes that are on top of the tree are the ones that contribute most to reducing the uncertainty about cluster membership, i.e. that usually the top attributes are the ones that most capture the “essence” of a cluster. We assume that this characteristic will naturally lead them to using attributes from the top of the tree when asked to provide a label for a cluster. At the same time, since a centroid representation does not provide any information about importance of attributes, we may assume that human analysts will choose attributes randomly or according to the order in which they are presented in the centroid. For rule representations, we may assume that an analyst will use those attributes that appear most frequently in the whole set of rules. These observations bring us to the following fourth hypothesis:

H4: When asked to provide a label to describe the essence of a cluster, a human analyst using

- a. a decision tree representation of the cluster will use attributes from the top of that tree in the label*
- b. a centroid representation will choose attributes randomly to formulate the label*
- c. a rule representation will use attributes in the label that appear most frequently within the rules*

Another assumption is that, when we explicitly ask a human analyst for the importance of an attribute that is at a low level of a decision tree representation, there is a risk that (s)he forgets about the instructions (see Sect. 4.2) and confirms that importance:

H5: Given a decision tree cluster representation and the question of whether an attribute at a low level of the tree is important for the essence of the cluster, a human analyst will be misled to say that the attribute is important.

The above hypotheses all aim at assessing the “effectiveness” aspect of representation, i.e. whether or not they lead a marketer to the right conclusions regarding the essence of a cluster. Regarding the efficiency aspect, i.e. the question how easy and fast marketers can derive the cluster essence using a particular representation, we need to formulate an open question (since we do not have an a priori assumption):

Q1: Given a set of alternative representations for a cluster, which representation will human analysts prefer? That is, for which representation will they state that it is easiest to derive the essence of a cluster?

And finally, we are interested to learn about the reasons for such preference:

Q2: Given the preference of a human analyst for a particular type of cluster representation, which reasons, i.e. which specific characteristics of the representation, lead to this preference?

4 Experimental Setup

4.1 Data Selection and Preparation

For our empirical validation of the hypotheses, we first needed a data set comprising description of persons by demographic and behavioural attributes. We chose the data of the 1998 KDD Cup¹, a data set that profiles persons, using 479 different attributes, with the purpose of predicting their reaction to a donation-raising campaign.

With that data, we proceeded as follows:

- We drew a random sample of 8000 persons
- We made a selection of 35 attributes, comprising 6 demographic attributes (combined socio-economic status and urbanicity level, age, home ownership, number of children, income and gender), 14 behavioural, numerical attributes indicating the number of known times a person has responded to other types of mail order offers and 15 binary variables reflecting donor interests (with values “yes” or “no”), as collected from third-party data sources.
- We clustered the data using k-means, setting the number of clusters to 15, and recorded the resulting centroids.
- We selected two clusters (cluster 5 and cluster 13) from that result.
- We built a decision tree (C4.5) and RIPPER rule model to distinguish between members of the two selected clusters and the rest of the data set, respectively.
- We represented the two decision trees in a user-friendly graphical form.

Thus, we had, for each of the two selected clusters, a centroid representation in a tabular format, a graphically represented tree and a set of text-based rules. Figure 1 shows the centroid representation of the two clusters – along with the mean and mode values for the full data set (which can be used for comparison and to detect potentially meaningful deviations). Figure 2 shows part of the tree representation and the full rule representation of cluster 5.

¹ <http://www.sigkdd.org/kdd-cup-1998-direct-marketing-profit-optimization>.

Attribute	Full Data	Cluster 5	Cluster 13
# persons	8000	79	818
neighbourhood	rural middle-class	urban upper-class	town middle-class
age	62.6855	53.1454	70.848
home owner	Y	Y	N
# children	1.5157	1.5125	1.5052
household income	3915.9	4303.8	2152.8
gender	F	F	F
Buy Craft articles	0	0	0
buy gardening articles	0	0	0
buy books	1	0	0
buy collectables	0	0	0
react to health pubs	0	0	0
buy male magazines	0	1	0
***	***	***	***
interest in collectables			
interest in veteran topics	Y		
interest in bible reading			
interest in catalog shopping	Y		
***	***	***	***

Fig. 1. Centroid representation (extract) of the two clusters

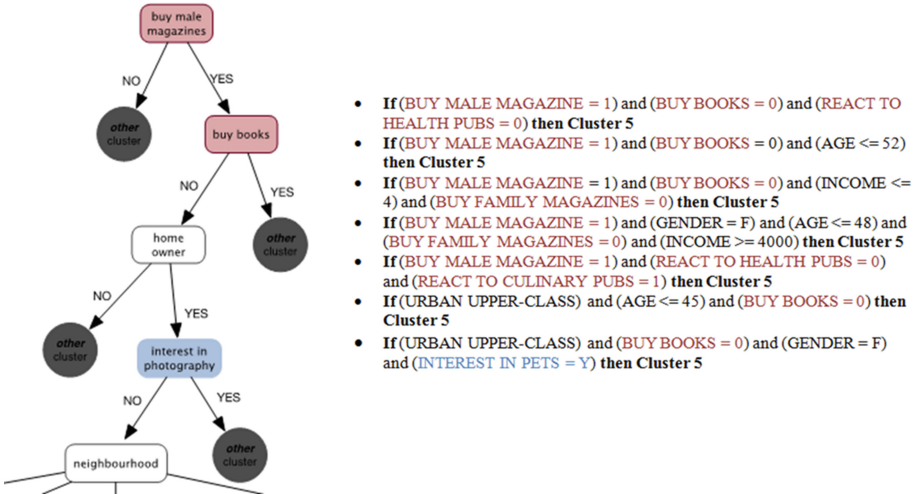


Fig. 2. Tree (extract) and rule representation of cluster 5

4.2 Participants and Questionnaire

Next, we prepared a set of tasks and questions to test the hypotheses presented in Sect. 3. We recruited a total of 46 participants who answered the questionnaire. We divided the participants into three groups, one for each type of representation, resulting in groups that we will call “centroid group”, “tree group” and “rules group”. Both the centroid and tree group had 15 members, the rules group had 16. All participants were students at our school. Hence, all of them

have come across the topic of statistics within their studies, i.e. should be familiar with statistical basics.

Each group received the same questionnaire, plus one sheet of paper containing the group-specific representation of both cluster 5 and 13 (e.g. a tree for the tree group). In addition, each participant was given a folded sheet containing the other two representations (e.g. rules and centroid for the tree group) of cluster 5. Participants were asked to leave that sheet folded, i.e. not to look at it before they were asked to. Finally, participants were instructed, both orally and in written form, about the meaning of attributes and the meaning of representations, e.g. it was mentioned that centroids contain modes and means, that decision trees have important attributes at the top and that rules may describe only a subset of a cluster.

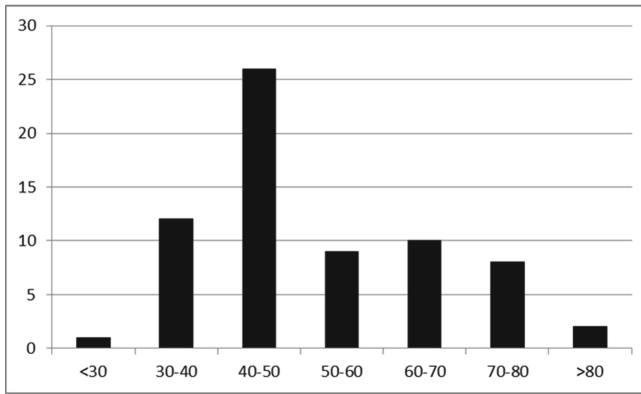


Fig. 3. Distribution of the age attribute in cluster 5

In the following, we report the questions of our questionnaire, along with our expectations regarding how each group of participants will react:

Task 1. To test hypothesis *H1*, we asked participants to look at their representation of cluster 5 and answer the question “What age group do you think most representative of this cluster?” Fig. 3 shows the age distribution within cluster 5 as a histogram – clearly the age group between 50 and 60 is not predominant in this cluster, as one could believe by analysing the mean value, 53.15 in the centroid (see Fig. 1), but we expect the centroid group to fall into that trap according to *H1*.

Task 2. For *H2*, we asked participants to look at their representation of cluster 13 and respond to the question “How would you make a generalisation about the demographic of this cluster by analysing the variables *neighbourhood*, *age*, *household income* and *gender*?”. Here, we are only interested in the “neighbourhood” attribute. When closely analysing its distribution, one finds that cluster members are not very rich (only 9% have the value “upper-class”). The other part of the attribute does not show a clear picture – the

mode of the attribute, which indicates that cluster members live in towns, only applies to 37 % of the cluster members. Hence, it is wrong to say that cluster members live in towns. Since the tree does not contain the neighbourhood attribute, we expect the tree group to indicate that no answer is possible and, according to $H2$, the centroid group to “fall into the trap” of saying “towns”. For the rule group – where various values of the attribute with various urbanicity levels are mentioned – we expect that only the socio-economic status (“middle or lower class”) will be mentioned by participants, but not the urbanicity level.

Task 3. To test $H3$, we let participants look at their representation of cluster 5 and then asked “Is there any indication of the health consciousness of this cluster and how would you describe it?” In cluster 5, the reaction to health pubs is more frequent (47 %) than in the full data (20 %). However, since two rules for cluster 5 contain the statement “REACT TO HEALTH PUBS = 0”, and according to $H3$, we expect that the rule group will say that cluster 5 members are predominantly not health conscious. For the other groups, we expect that they do not see an indication since the tree does not contain the attribute and the centroid does not report a value that deviates from the full data.

Task 4. For $H4$, we asked participants to write down a short label to characterise the essence of cluster 5. We then analysed which attributes they had used in their labels and compared them to our expectations according to hypotheses $H4$, a), b) and c).

Task 5. To test $H5$, we asked participants to infer, from their representation of cluster 13, an answer to the question “Is pet ownership a significant characteristic of this cluster?” According to $H5$, we expected all groups to deny this importance.

Task 6. Finally, we told participants to open the folded sheet containing the alternative representations of cluster 5 and then asked the question “Would you have found it easier to label the clusters with one of these alternative representations? If so, which of the two?” From the answers, we collected evidence to answer our research questions $Q1$ and $Q2$.

5 Results and Discussion

Since all questions that we gave to the participants were to be answered freely, i.e. not using multiple choice, their answers had to be coded before a quantitative analysis became possible. Below, we report such analysis and discuss the results, at the same time explaining the codes that we formed inductively while reading the answers. Sometimes, coding required interpretation, especially when analysing the cluster labels that participants were asked to write down in Task 4. For instance, the tree representation of cluster 5 does not contain the gender attribute. But it does contain the attribute “buy male magazines” at the top of the tree (see Fig. 2). Hence, when participants stated that cluster members were male, we assumed that they used the “buy male magazines” attribute to infer that.

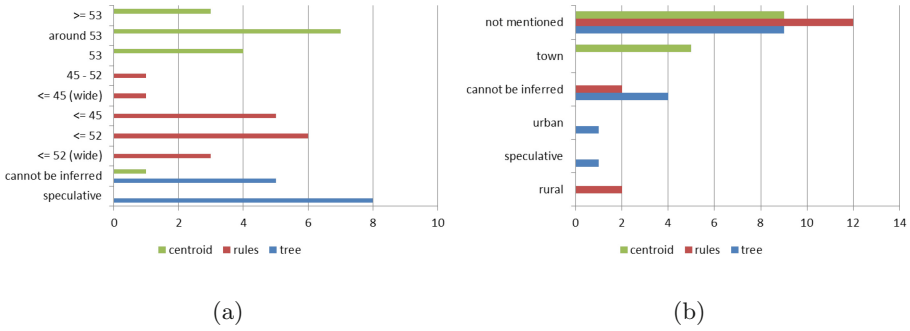


Fig. 4. Frequency of answers (coded) for Tasks 1 and 2 (colour in on-line proceedings)

Figure 4(a) shows the coded answers and their frequency of occurrence for Task 1. Here and henceforth, the code *cannot be inferred* means that a participant made an explicit statement that there is no or not enough evidence to infer an answer to the task. We generally use the code *speculative* to summarise all answers where the participant gave an answer that did not refer to values of the attribute in question, but where it was obvious that this value was inferred from other attributes. For instance, a member of the tree group gave the answer “15 to 25 (income less than 4000) or 65 to 80 (income lower, home owners, collectors, gardening)” for Task 1, which shows that age was inferred from income and some interest variables. The codes ≤ 45 (wide) or ≤ 52 (wide) summarise all answers where the participants indicated that *all* ages below 45 and 52 respectively are included. The frequencies of answers show very clearly that all members of the centroid group fell into our trap and derived an answer from the (misleading) mean. Only one centroid group member indicated that “cluster 5 only has the average age”, which we coded as *cannot be inferred*. Members of the tree group either speculated or indicated that age cannot be inferred since the age attribute is not present in the tree. Members of the rule group made statements that better reflect the actual age distribution as shown in Fig. 3. In summary, we conclude that we can clearly accept hypothesis $H1$.

Figure 4(b) shows coded answers and their frequency for Task 2. We only coded the urbanicity level part of the answer, which was sometimes not mentioned in the answer (resulting in the code *not mentioned*). The codes *cannot be inferred* and *speculative* are defined as above. We can see that, if members of the centroid group mention the urbanicity level, they say “town”. Only three members of both other groups mention the urbanicity level at all, consistent with our expectations. Thus, although the support for this is smaller than expected (with only 5 explicit answers from the centroid group), we can carefully confirm $H2$.

Next we look at the result of Task 3, displayed in Fig. 5(a). Here, the codes *yes, positive* and *yes, negative* denote that participants said that there is an indication of health consciousness and that it is positive or negative, respectively. The code *yes, undirected* means that the participants only said there is indication of health consciousness but did not indicate anything from which to conclude the

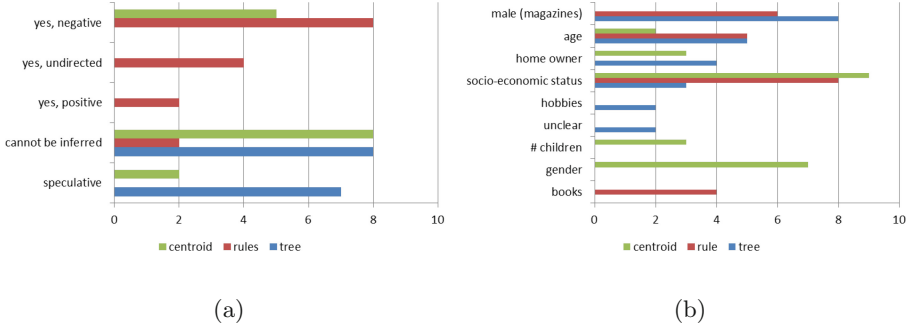


Fig. 5. Frequency of answers (coded) for Tasks 3 and 4 (colour in on-line proceedings)

direction. The results show that, as expected, 8 members of the rule group fall into the trap of concluding, from statements embedded in only two rules, that cluster 5 members are not health conscious. We can assume that if we insisted on an answer to the direction question, some of those 4 rule group members who indicated no direction would also choose a negative health consciousness. Hence, we can conclude that there is at least a rather large danger of such conclusion for human analysts with a rule representation and thus carefully confirm *H3*. The danger also seems to exist – to a smaller extent – for the centroid group: although the mode here does not deviate from the full data, some people are led to believe that health consciousness is negative, just because the value is 0 for this attribute.

The results of Task 4 are displayed in Fig. 5(b). Besides the names of attributes, we have used the codes *unclear*, which means that no direct reference to an attribute could be detected in the answer and *hobbies* which describes answers that refer to all interest attributes as a whole. Here, the sum of frequencies of codes across each group are larger than the group size since many participants used more than one attribute in their answer. The results show that the different groups use rather different attributes for labeling. We see that there is a strong tendency for members of the tree group to use attributes from the top of the tree, i.e. mainly “buy male magazines” and “home owner”. It is a bit surprising that “buy books” is not used although at second-highest position in the tree and that many participants in this group use age which can only be speculative since the tree does not contain the age attribute. We may assume that having carried out Task 1 before had an influence here – especially since the same people who speculated in Task 1 did the same here and their answers were consistent between Tasks 1 and 4. When inspecting the answers from the centroid group, we see that participants used exclusively demographic attributes to describe the cluster – which may be explained by the fact that they are at the top of the centroid representation (see Fig. 1). Their choice is hence, although not completely random, not motivated by any meaningful criteria. Members of the rule group use primarily the attributes *socio-economic status*, *buy male*

magazines, *age* and *buy books*. Except *age*, all of these are frequent or mid-frequent in the rules for cluster 5, see Fig. 2. All of these observations, taken together, let us rather clearly accept *H4*.

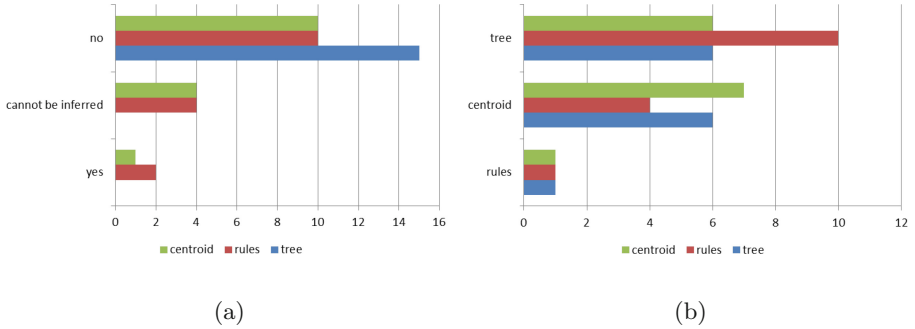


Fig. 6. Frequency of answers (coded) for Tasks 5 and 6 (colour in on-line proceedings)

Figure 6(a) shows answers for Task 5. We can see that *H5* can be rejected very quickly since no member of the tree group found pet ownership to be a significant characteristic of cluster 13. Unexpectedly, a total of three participants from both the centroid and rules group think that this is the case – for unknown reasons.

Finally, we see the preferences of participants regarding the representation that allows labeling with greatest ease in Fig. 6(b). We see that rules are unpopular in all groups. Trees and centroids are equally popular in the tree and centroid groups. Participants from the rules group show a stronger preference for the tree such that this is also the overall most popular choice. Finally, we have coded the reasons that participants gave to explain their preference (if any). The code *simplicity* refers to answers that said that a representation is easy to understand/grasp. *Completeness* means that participants liked the fact that a representation gives overview over all (or enough) attributes. The attribute *comparison* refers to the ease of comparing several cluster, *attribute order* to the ease of identifying important attributes easily and *precision* to the space that a representation leaves for false interpretations.

Arguments in favour of the tree representation were mainly simplicity (7 mentions), as well as attribute order and precision (1 mention each). Rules – if preferred – were liked for their *completeness* (2 mentions). The centroids are also preferred because of *completeness* (3 mentions), and additionally for their *simplicity* and possibility of *comparison* (1 mention each). To sum up, the main reason for choosing a tree – according to our participants – is *simplicity*, whereas a possible reason for choosing centroids could be their *completeness*.

6 Conclusions and Future Work

In our experiments, we were able to confirm all of our intuitions about the process of capturing cluster essence and possible false conclusions that may result. In

summary, we saw that centroids have several severe weaknesses, both in the sense that it is hard to identify the attributes that most contribute to cluster essence and that false conclusions may result from looking only at mean or mode values – even for humans who have a moderate background in statistics.

Decision trees do not show such undesired characteristics and are also most popular with our participants when it comes to the question of how easily cluster essence can be inferred from a representation. Our experiment has also revealed additional interesting arguments in favour of certain representations, which need to be assessed in light of our other results as follows: in light of the traps that people fell into, the only really valid argument in favour of centroids is their strength in allowing to easily compare several clusters. Trees score in simplicity. Attribute order – again considering the confirmation of our hypotheses – should also be seen as an important argument in favour of decision trees since it results in labels that better reflect truly important attributes. In this context, completeness should not be counted as a valid argument (used in favour of centroids or rules) since it does not serve a meaningful purpose to show summaries of attributes that do not contribute to describing the essence of a cluster – and may only lead to false conclusions about that essence.

For future research, it will be interesting to investigate more possible cluster representations, such as ones resulting from the application of expectation maximisation or COBWEB clustering – and to develop and test new hypotheses that go along with those representations. This includes also advanced centroid representations that use e.g. confidence intervals for means, try to order attributes by importance or show full distributions of categorical attributes. Similarly, future work might want to study improved rule representations, e.g. by showing the coverage of rules or play with different levels of pruning the decision trees. In addition, to get a deeper understanding, one might perform separate in-depth analyses for each kind of attribute type, study different kinds of distributions of numerical attributes and analyse the effects of visualisation techniques.

References

1. Cohen, W.: Fast effective rule induction. In: Proceedings of the Twelfth International Conference on Machine Learning, pp. 115–123 (1995)
2. Dempster, A., Laird, N., Rubin, D.: Maximum likelihood from incomplete data via the EM algorithm. *J. R. Stat. Soc. Ser. B (Methodol.)* **39**, 1–38 (1977)
3. Radev, D.R., Jing, H., Stys, M., Tam, D.: Centroid-based summarization of multiple documents. *Inf. Process. Manage.* **40**(6), 919–938 (2004)
4. Fisher, D.: Knowledge acquisition via incremental conceptual clustering. *Mach. Learn.* **2**(2), 139–172 (1987)
5. Gordon, A.: Classification, 2nd edn. Taylor and Francis, London (1999)
6. Huang, Z.: Clustering large data sets with mixed numeric and categorical values. In: Proceedings of the First Pacific-Asia Conference on Knowledge Discovery and Data Mining, pp. 21–34 (1997)
7. Jain, A.K.: Data clustering: 50 years beyond K-means. *Pattern Recogn. Lett.* **31**(8), 651–666 (2010)

8. Kotsiantis, S.: Supervised machine learning: a review of classification techniques. *Informatica* **31**(3), 249–268 (2007)
9. MacQueen, J.: Some methods for classification and analysis of multivariate observations. In: *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, vol. 1, pp. 281–297 (1967)
10. Popescul, A., Ungar, L.: Automatic labeling of document clusters (2000), <http://citeseer.nj.nec.com/popescul00automatic.html>
11. Quinlan, R.: *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers, San Mateo (1993)
12. Saglam, B., Salman, F., Sayin, S., Trkay, M.: A mixed-integer programming approach to the clustering problem with an application in customer segmentation. *Eur. J. Oper. Res.* **173**(3), 866–879 (2006)
13. Teichert, T., Shehu, E., von Wartburg, I.: Customer segmentation revisited: The case of the airline industry. *Transp. Res. Part A: Policy Pract.* **42**(1), 227–242 (2008)
14. Treeratpituk, P., Callan, J.: Automatically labeling hierarchical clusters. In: *Proceedings of the 2006 International Conference on Digital Government Research*, pp. 167–176 (2006)
15. Weber, R.: Customer segmentation for banks and insurance groups with fuzzy clustering techniques. In: *Fuzzy Logic*, pp. 187–196 (1996)
16. Wedel, M., Kamakura, W.: *Market Segmentation: Conceptual and Methodological Foundations*, 2nd edn. Kluwer Academic Publishers, Boston (2000)

Advances in Data Mining: Applications and Theoretical
Aspects

15th Industrial Conference, ICDM 2015, Hamburg,
Germany, July 11–24, 2015. Proceedings
Perner, P. (Ed.)

2015, X, 279 p. 65 illus., Softcover

ISBN: 978-3-319-20909-8