Application of Data Mining Techniques to the Data Analyses to Ensure Safety of Medicine Usage

Masaomi Kimura
Shibaura Institute of Technology
Japan

1. Introduction

Not only the safety of medicines themselves, from the perspective of medicinal properties, but also the safety of medicine usage is important, in order to ensure the right use of the right medicines and prevent medical accidents. The author has applied data-mining techniques to the analyses of data to investigate the status about the safety of medicine usage, such as the data in the database of medical near-miss cases which have been collected by Japanese government, the investigation data to understand how patients handle the injection device for anti-diabetic drug by themselves after the guidance by a doctor, the questionnaire data asking opinions about the mark indicating a medical property on some cardiac transdermal patch, which were provided by medical experts and/or pharmaceutical companies.

The analyses on such data have been traditionally based on the statistical approaches, which bring us the information about the single attribute of data, for instance, the tables showing the frequency of events under some condition and the analysis whether the difference of the frequency is statistically significant, and so on. It is, however, important to extract the information not only of the each attribute independently but also about the relations among them, since each data in the results of the investigations is supposed to be generated under complex conditions, some set of which is an essential cause. In order to find the relationships of data or the items of the data, it is useful to utilize the clustering algorithms to classify the data into some clusters, in each of which data have the same characteristics, and the decision tree algorithms to find the condition to distinguish the data based on the classes to which they belong, from in rough to in detail. By means of both algorithms, we can find a structure among the data, which kinds of data we have obtained and which conditions generate such kinds of data.

In this paper, we first make a brief review of some clustering algorithms including agglomerated hierarchical clustering algorithm, TwoStep clustering algorithm and K-means algorithm with the number of seeds optimized by Bayesian information criterion (BIC) and also show how to combine them to the idea of specialization coefficients. Next, we introduce the review of the application to the data related to the safety of medicine usage shown above. Last, we summarize the principle of application of data mining to such analyses.

Source: Data Mining and Knowledge Discovery in Real Life Applications, Book edited by: Julio Ponce and Adem Karahoca,
ISBN 978-3-902613-53-0, pp. 438, February 2009, I-Tech, Vienna, Austria

2. Clustering algorithms and their variations for applications to questionnaire data / survey data

As we stated in the introduction, clustering algorithms help us find the structures that lie within data(Berry & Linoff, 1997). There are two types of algorithms, hierarchical one and non-hierarchical one. The representative algorithm of the former type is the agglomerated hierarchical clustering algorithm, which incorporates data or clusters sitting neighbourhood in the order of short distance. This provides us with graphical output, dendrogram, which shows the global structure of data in the form of a tree diagram. When we apply this method, it is important to adopt appropriate distance measure between data and/or clusters. The representative algorithm of the latter type is K-means algorithm, which classifies data into the predetermined number of clusters and shows masses within data. Though this offers easy-to-understand results, unfortunately, it is not clear which number of cluster is appropriate to adopt in the typical way of application. Another typical algorithm, which is the intermediate algorithm of hierarchical and non-hierarchical one, is TwoStep clustering algorithm (SPSS inc., 2003; Zhang et al., 1996), which can be effectively applicable to massive data. This consists of two procedures. To deal with data effectively, the data is not handled homogeneously but divided into groups (pre-clusters) that consist of similar data by means of CF Tree (the first step). After that, clustering algorithm is again applied to the pre-clusters (the second step).

In order to apply such algorithms to questionnaire data or survey data, we propose some improvements in following subsections.

2.1 Agglomerated hierarchical clustering using selection co-occurrence measure

Let us consider the case to deal with multiple-choice questions. In general, the analysis of the answers for such questions is based on the majority vote for each option. However, the combinations of the selected options also have a significant meaning, since the answerers do not choose each option independently but they express their intent as a group of options. Hence, it is important to find the option patterns in the data that most answerers select, by utilizing clustering algorithms.

Let d denote the number of options, and n, the number of respondents. Let us also define 'answer matrix' A_{ij}, whose value is '1' if the ith respondent selects the jth option and '0' if not. To see the co-occurrence relationship between the selection of options, we apply an agglomerated hierarchical clustering algorithm to the column vector of the answer matrix, which we call selection vector $A_{i*} = (A_{i1}, A_{i2} \ldots A_{in})$.

When we perform agglomerated hierarchical clustering algorithm, we usually adopt the Euclidean distance or Manhattan distance to measure the distance of the vectors. If we apply them to two selection vectors, it is equivalent to counting the number of different elements of the vectors, or counting the number of '1' obtained from the result of 'exclusive or' of each corresponding element of the vectors. This results in not only the most simultaneously selected options, but also simultaneously unselected options being judged to neighbour each other. The options that are unselected by most respondents therefore have a short distance to each other.

The similarity index, however, should measure the co-occurrence of simultaneously selected options, not of those unselected. We therefore adopt other similarity measures than the Euclidean or Manhattan distance.

Let us remember the definition of the inner product of the vectors. If the elements of the vector are either '1' or '0', the inner product counts up the number of elements whose value is both '1', in other words, the number of '1' obtained from the result of the 'and' operation to each corresponding element of vectors. The contributing value comes only from the elements that the respondents choose as a cluster. This suggests us that the inner product is suitable for the similarity index. We take this into account and propose the 'distance' that counts the number of elements whose value is not 1 in either vector. Let $D^{\#}$ denote the 'distance', which can be calculated as

$$D^{\#}(A_{i*}, A_{j*}) = \sum_{k=1}^{n}(1 - A_{ik}A_{jk}) = n - \sum_{k=1}^{n} A_{ik}A_{jk} \tag{1}$$

Note that $D^{\#}$ does not satisfy one of the axioms of distance, namely $D^{\#}(A_{i*}, A_{i*})$ is not equal to 0. This is because the contribution to $D^{\#}$ of the element that has the value 0 in each vector is defined as 1. The requirement of distance for agglomerated hierarchical clustering is to supply measurement that allows us to compare the similarity. Since measure $D^{\#}$ satisfies the requirement, we adopt $D^{\#}$ as 'distance'. After this, we call $D^{\#}$ the selection co-occurrence measure (Kimura et al., 2006a).

As merging methods of clusters, many methods are known, such as the nearest-neighbour method, the furthest-neighbour method and the Ward method. We applied these methods and found that they provide similar dendrograms as a result. Not some but all of the options in the clusters should have a similar pattern of selection if they are merged. We therefore adopt the furthest-neighbour method as the clustering algorithm.

Table 1 shows the sample data that we used to verify our method. Figure 1 shows the dendrogram for which the Euclidean distance is used and that for which we use the selection co-occurrence measure. We can see that Option 2 and Option 5 are judged to be neighbouring for Euclidian distance, although no respondents chose them simultaneously. However, for selection co-occurrence measure, these options are judged to be far from each other as we expected.

	Option 1	Option 2	Option 3	Option 4	Option 5
Answerer1	1	0	1	1	0
Answerer2	1	0	1	1	0
Answerer3	1	0	1	1	0
Answerer4	0	0	1	1	1
Answerer5	0	0	0	0	1
Answerer6	0	1	0	0	0
Answerer7	1	1	1	1	0
Answerer8	1	0	1	0	0
Answerer9	1	0	1	0	0

Table 1. Sample data to verify selection co-occurrence measure.

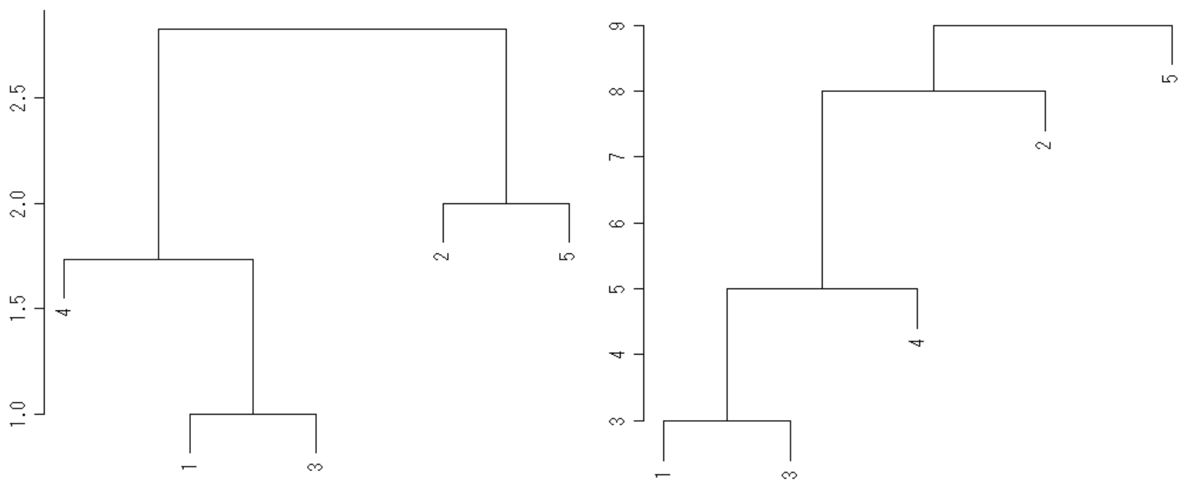


Fig. 1. Dendrogram for Euclidian distance (left) and selection co-occurrence measure (right).

3.2 K-means algorithm with the number of clusters determined by BIC

The K-means algorithm classifies data into clusters (the group of data that neighbor each other), the number of which is specified beforehand. (Let K denote the number of clusters.) The algorithm is realized by the iteration of the following procedures:

A. Centroid X_i is obtained as Equation 2, where C_i denotes the ith cluster and N_i denotes the number of elements of cluster C_i.

$$X_i = \frac{1}{N_i} \sum_{x \in C_i} x \tag{2}$$

B. Measuring the distance from the data to each centroid, the data is reassigned to the i_xth cluster specified as Equation 3.

$$i_x = \arg\min_i \left\| x - X_i \right\| \tag{3}$$

We adopt converged clusters after the iteration of Step A and B. Note that we do not have C_i initially in Step A, the randomly selected data are used as substitutes.

As we mentioned above, the number of clusters K has to be specified before performing the algorism and is usually determined by trial and error. In this study, we determined it by utilizing Bayesian information criteria (BIC). Although it should be simple for the model to suppress overfitting to the data, there is a trade-off relationship between the simplicity and the accuracy of the model. BIC is the index that balances the simplicity and the accuracy of model, whose minimum value gives the best model for describing the data.

To calculate the value of BIC, we translate the K-means method into a probability model, or more specifically, a likelihood function. We assume the likelihood function of the K-means method as follows:

$$p\left(\{x\}\right) = \prod_{i=1}^{K} \prod_{x \in C_i} P_i(x; X_i, \sigma) \tag{4}$$

$$P_i(x; X_i, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{\|x - X_i\|^2}{2\sigma^2} \right) \tag{5}$$

where {x} denotes the data group that has n elements, and X_i and σ are parameters whose maximum likelihood estimates are the centroid of cluster C_i and the mean value of the variance in each cluster. With these equations, the K-means algorithm can be reproduced as follows.

A'. Calculate the maximum likelihood estimates of X_i and σ of *p({x})* as

$$\frac{\partial}{\partial X_i} p\left(\{x\}\right) = 0 \tag{6}$$

$$\frac{\partial}{\partial \sigma^2} p\left(\{x\}\right) = 0 \tag{7}$$

As a result, we obtain the centroid of each cluster.

B'. Each item of data is assigned to the cluster that has the maximum value of probability P_i as is shown in Equation 8. In fact, this condition is equivalent to Equation 2, since exponential function is monotonic.

$$i_x = \arg\max_i P_i(x; X_i, \sigma) \tag{8}$$

We again obtain the same clusters that is obtained by iterations A) and B), since steps A′) and B′) are equivalent to A) and B), respectively. We therefore adopt *p({x})* as a likelihood function that corresponds to the K-means algorithm.

The definition of BIC is given as Equation 9 (Shimodaira et al., 2004):

$$BIC = -2\sum_x \ln P(x; \hat{\theta}) + F \ln n \tag{9}$$

where the first term is the logarithm of the likelihood function with the maximum likelihood estimates, F is the degree of freedom of the parameters and n is the number of data. The first term decreases its value if we make many small clusters that fit the data. The second term, however, decreases its value if we make a small number of clusters. The smallest value of BIC gives us the optimal value of K, which provides a small number of clusters approximating the data well. The model that we use in our study has K centroids and one variance as parameters. If the dimension of the data is D, the degree of freedom is KD+1. BIC for our model (Kimura et al., 2006b) is therefore given as:

$$BIC = n\left(1 + \ln\left(\frac{2\pi}{n} \sum_{i=1}^{K} \sum_{x \in C_i} \|x - X_i\|^2 \right) \right) + (KD+1)\ln n \tag{10}$$

For clustering, we calculate BIC for the target data and find K that minimizes BIC.

3. Application

We review our studies where the methods shown in the previous section are applied. First, we show the application of agglomerated hierarchical clustering with selection co-occurrence measure to the questionnaire data that investigate the evaluation of design of 'therapeutic classification mark' of cardiac transdermal patch. (Kimura et al., 2006a) Next, we introduce the application of BIC to K-means algorithm to evaluate the characteristics of patients who handle the injection device for anti-diabetic drugs in the wrong way even after guidance. (Kimura et al., 2006b)

3.1 The application of agglomerated hierarchical clustering with selection co-occurrence measure

Our target data was the answers for the multiple-choice questions listed in Table 2, which is a part of the questionnaire on the 'therapeutic classification mark' and product name label to ensure the safety of drug use, performed in 2004. Each question is answered by 7,078 doctors, 7,018 nurses and 7,361 pharmacists.

It was important to conduct and analyze such nation-wide investigations using questionnaire to obtain feedback from medical experts (doctors, nurses, pharmacists) and patients about the 'therapeutic classification mark' printed on isosorbide dinitrate transdermal patches, which is a cardiac medicine, since there had never been a study that estimates the measures from the position of medical experts or patients. We compare the result obtained by agglomerated hierarchical clustering algorithm with selection co-occurrence measure with the one obtained by TwoStep clustering algorithm in order to see what groups of respondents select the patterns (combinations) of options. Figure2 shows the result that is obtained by applying the method to the sample data in Table 1. Roughly speaking, this suggests that there are two clusters, one of which contains Options 1, 3 and 4 and the other of which contains Option 2 and 5.

Comparing this with Fig. 1, we can see the relationship among the results as follows:

- Although Option 4 is found to be a neighbor of the cluster of Option 1 and 3 in Fig. 1, it is difficult to determine whether we should regard these three options as one group. Figure 2, however, suggests that the options can be identified as one group.
- Figure 2 suggests that Option 2 and Option 5 are chosen simultaneously, although they have never actually been chosen together. This is because each of the options is independently selected with Options 3 and 4, which mediate Options 2 and 5 to be in the same cluster. Figure 1, however, shows that Option 2, and 5 cannot be included in the same cluster.

Considering these results together, we can identify the group of options that the respondents select together, by comparison with these two results. On this point, these two methods complement each other.

Figure 3 shows the results of Question A. We can see that Option 1 and 3 are selected together. Although the frequency of co-occurrence is smaller, Option 5 can be regarded as being selected simultaneously with those two options. This suggests that the combination of systemic transdermal absorbent preparations that medical experts deal with is mainly a cardiac drug and an asthma drug, and that there are some cases where cancer pain-relief medicine is also used.

The results for Question B are shown in Fig. 4. From these results, we can see that Option 1 and Option 3 are selected together by most respondents, and Option 5 is the one that is simultaneously selected next to these options. This suggests that the main reason to adopt systemic transdermal absorbent preparations as a dosage form is that no burden is imposed on the digestive tract and that the effect continues for many hours. Additional to this, it is also because diet does not have any impact. Since we can also see that the selection of

Options 2, 4 and 6 are associated with these options, the effect of liver, administration termination and good compliance can be regarded as associated reasons to select transdermal patches.

A		***What systemic transdermal absorbent preparations do you usually deal with?***
	1	Cardiac medicine
	2	Hormone replacement
	3	Asthma medicine
	4	Smoking-cessation medicine
	5	Cancer pain-relief medicine
B		***Why did you select the systemic transdermal absorbent preparation?***
	1	Burden is not imposed on the digestive tract.
	2	First pass effect of the liver does not have an effect.
	3	Effect lasts for many hours.
	4	Administration can be terminated by peeling off.
	5	Eating meals does not have an effect.
	6	I can ensure good compliance.
	7	I do not select.
	8	Others
C		***What do you think about the design of the therapeutic classification mark and product name label of Frandol tape S?***
	1	The concept is valid for medical accident prevention.
	2	More innovation of the concept is necessary to prevent medical accidents.
	3	The mark is favourable for cardiac transdermal patches.
	4	More innovation of the mark is necessary.
	5	The print colour, white, is easy to see and favourable.
	6	The print colour should be more vivid.
	7	The mark, label and layout are valid for medical accident prevention.
	8	More innovation of the size of the mark, the number of labels, and layout is necessary.
D		***What are preventive measures against medical accidents related to the systemic transdermal absorbent preparation?***
	1	Displaying the mark and the label is good enough.
	2	The mark for the same efficacy should be integrated.
	3	The mark for the same efficacy should not be integrated but unique to each company.
	4	The mark should be displayed for other systemic transdermal absorbent preparations.
	5	Displaying the mark and the label is unnecessary.
	6	The effort is necessary to earn recognition from medical experts, patients and their families.

Table 2. Questions and options (originally in Japanese)

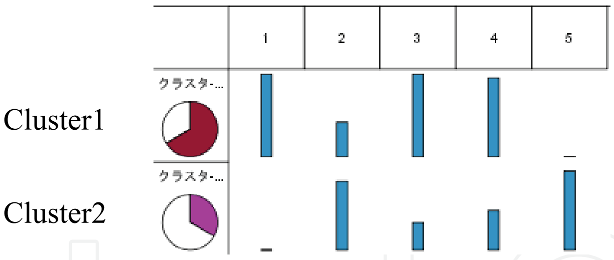


Fig. 2. The result of TwoStep algorithm applied to the sample data in Table 1.

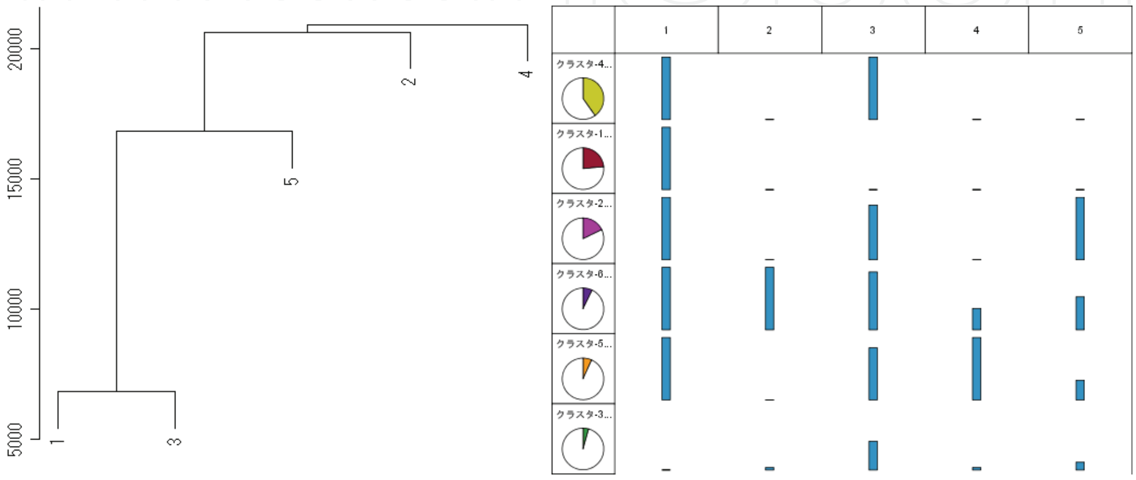


Fig. 3. The dendrogram (left) and clusters obtained by TwoStep algorithm (right) of selected options of Question A.

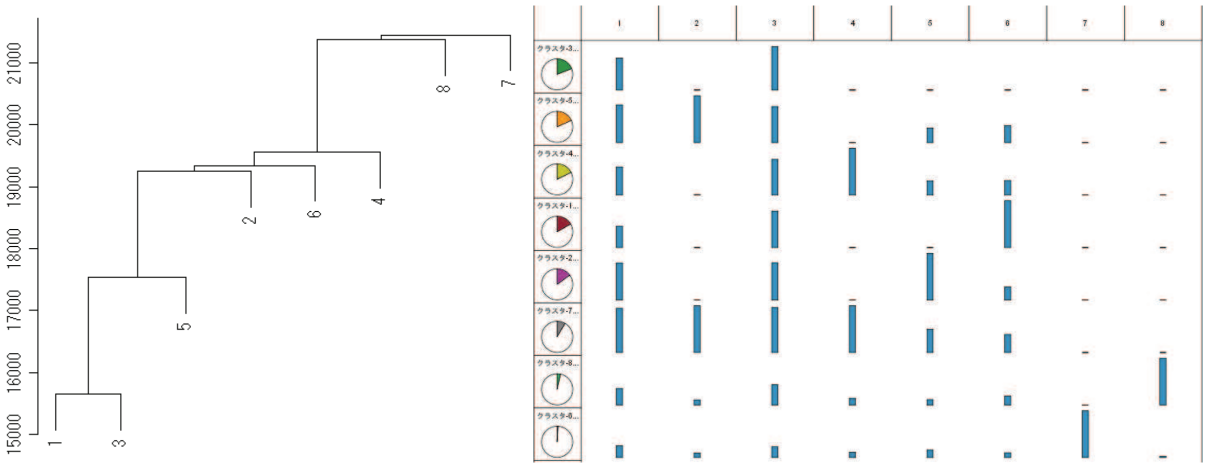


Fig. 4. The dendrogram (left) and clusters obtained by TwoStep algorithm (right) of selected options of Question B.

The results of Question C are indicated in Fig. 5. Comparing these figures, we can see that Options 1, 3, 5 and 7 are selected together. This can be interpreted as the design being valid for the prevention of medical accidents caused by transdermal patches, having a high level of visibility and being preferable. Since the cluster that includes 1, 3, 5, and 7 in the result of TwoStep algorithm contains 81% of the respondents, this suggests that most medical experts have a favorable opinion on the mark and the product name label on the target transdermal patch. Though some reader might think only 5,000 respondents select the pair of Options 1 and 3 in the dendrogram, they should notice that the result of TwoStep algorithm counts the number of respondents who chose any combination of Options 1, 3, 5 and 7.

Figure 6 shows the results of Question D. These figures indicate that most respondents selected Options 2, 4 and 6. This suggests that most medical professionals think that the therapeutic classification mark and product name label are necessary, should be integrated for the same efficacy and should be widely recognized. The respondents in the cluster in Fig. 6 account for about 58% of the whole. This indicates that more than half of the respondents do not satisfy the current situation and think that use of the mark and product name label should be widely spread.

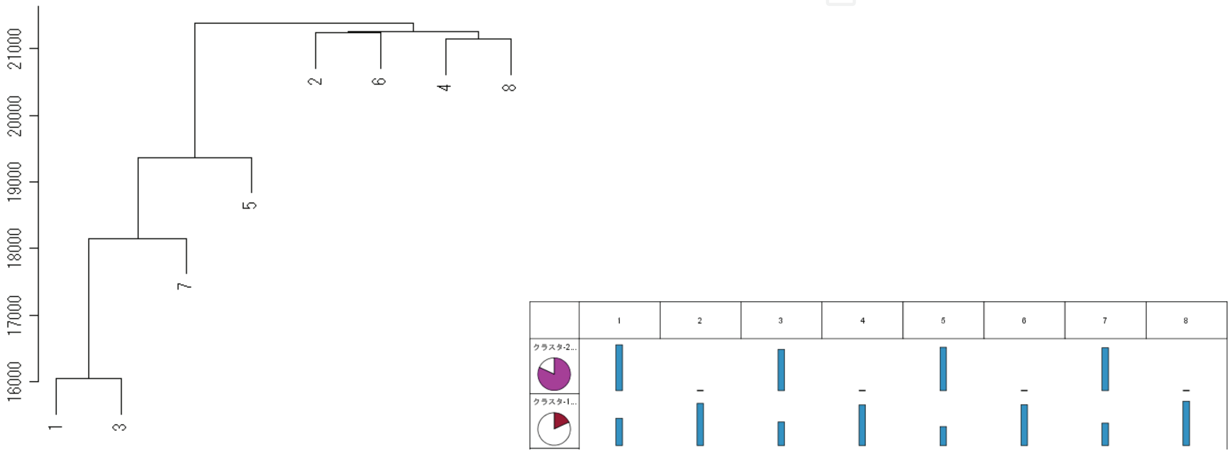


Fig. 5. The dendrogram (left) and clusters obtained by TwoStep algorithm (right) of selected options of Question C.

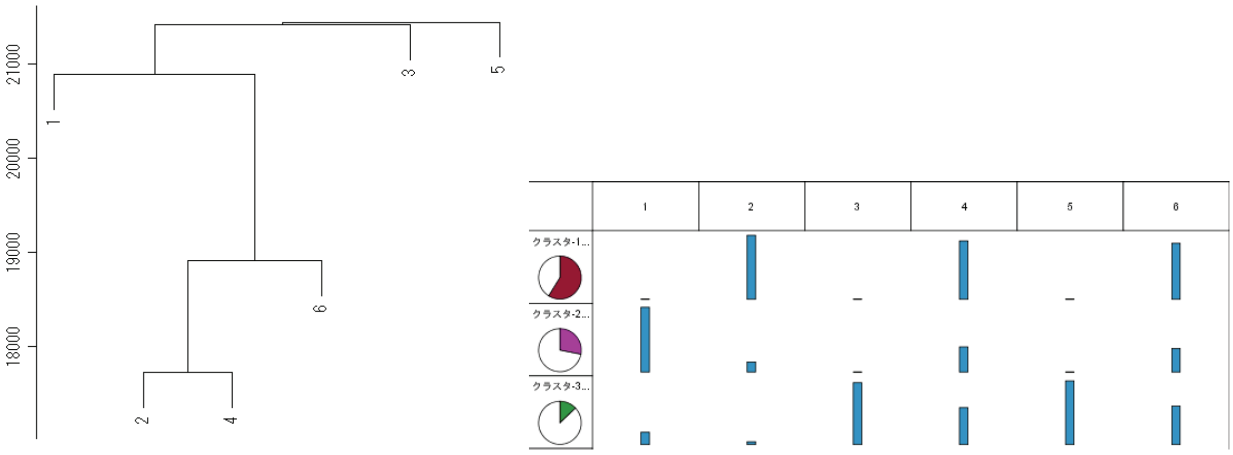
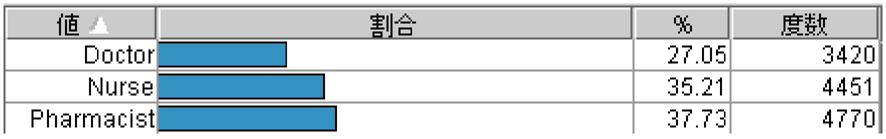
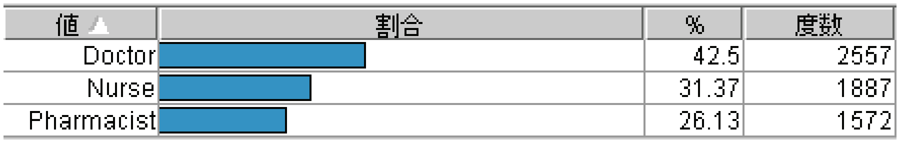


Fig. 6. The dendrogram (left) and clusters obtained by TwoStep algorithm (right) of selected options of Question D.



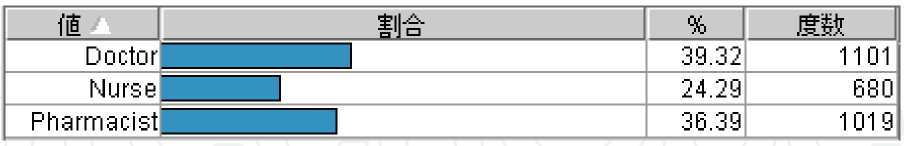
値	割合	%	度数
Doctor		27.05	3420
Nurse		35.21	4451
Pharmacist		37.73	4770

Fig. 7. The distribution of occupation in Cluster 1.



値 ▲	割合	%	度数
Doctor		42.5	2557
Nurse		31.37	1887
Pharmacist		26.13	1572

Fig. 8. The distribution of occupation in Cluster 2.



値 ▲	割合	%	度数
Doctor		39.32	1101
Nurse		24.29	680
Pharmacist		36.39	1019

Fig. 9. The distribution of occupation in Cluster 3.

Figures 7, 8, and 9 show the distribution of respondents in each cluster in 12 by occupation. From Fig. 7, we can see the trend that the opinion that the mark and label should be promoted comes from pharmacists and nurses. Figure 8 suggests that doctors tend to satisfy the current situation. Figure 9 shows that the profession people who least select the negative options are nurses.

3.2 The application of K-means algorithm with the number of clusters determined by BIC

In this subsection, we review the study whose target data was a self-injection device of an antidiabetic drug, insulin. Before using the device, patients need to conduct procedures such as setting the medicinal solution cartridge and the injection needle, confirmation by pulling the cartridge to ensure secure attachment and test injection.

In this procedure, it is found that 'confirmation by pulling the cartridge system before setting the units of the test injection', has the lowest success rate. In the procedure, patients confirm that the cartridge is firmly mounted to the device in the trial test injection. A test injection is important to prevent the mixture of air in the injection device. If patients do not perform the confirmation procedure in the right way, it is not ensured that test injection has been correctly performed. In this study, we classify the type of examinees who could not accomplish the confirmation procedure by four parameters: age, length of use of the pre-improved device, length of use of a pen-type insulin injection device, and length of supervision by medical experts.

The investigation consists of two trials that are performed at certain intervals for the same examinee. The total number of examinees is 589, the number of failed examinees in the first trial is 199 and the number of failed examinees in the second trial is 264. We also compare the results of the trials.

In this study, we utilize K-means algorithm with the number of clusters determined by BIC and obtain knowledge on the tendency of patients who cannot accomplish the procedure. After performing K-means clustering, we compare the distribution of the failed examinees in each trial as mentioned above. To do this, we utilize the specialization coefficient, which is obtained as follows:

a. Calculate the relative frequency of the number of examinees in each cluster who could not accomplish the procedure.
b. Calculate the relative frequency of the number of all examinees in each cluster.
c. Obtain the ratio of the value of a) to b).

The reason that we use the ratio of relative frequency rather than the ratio of the number of examinees is that we can normalize the ratio so as to let it be 1 if the distribution of the failed examinees is the same as the distribution of all examinees for each cluster.
To compare the tendency of the failed examinees in detail, we also obtain the specialization coefficient of the groups, which is classified by cluster, gender and age.
We calculate BIC for K whose value is between 1 and 8 (Fig. 10) and found the minimum value of BIC is given by K=3 for our target data. We therefore applied this value to the K-means algorithm for the profile data of examinees that fail to accomplish the operation in the first trial.

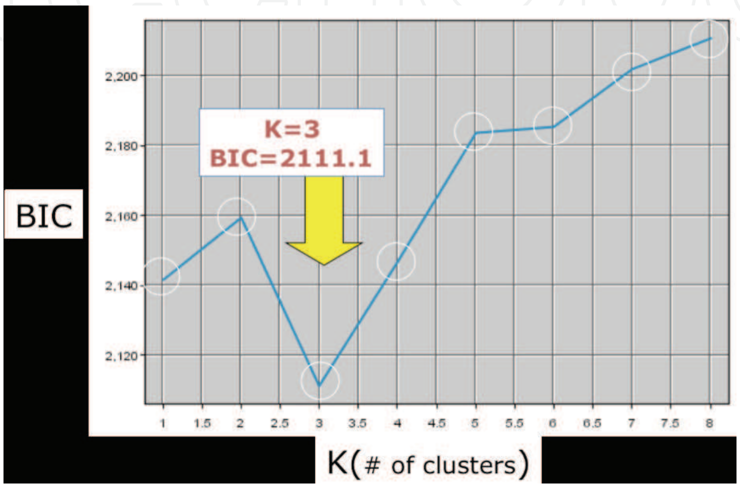


Fig. 10. The relationship between BIC and K (the number of clusters) for the target data.

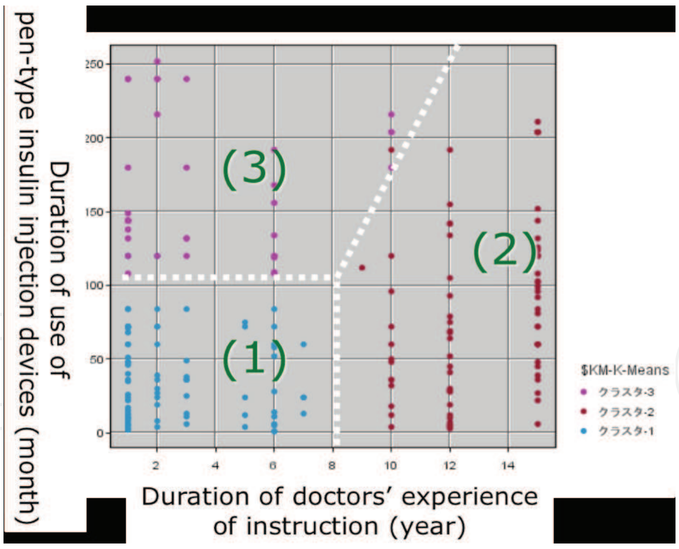


Fig. 11. The resultant clusters.

The result of the K-means algorithm is shown in Figure 11. From this, we can see that two parameters, the length of use of a pen-like insulin injection device and the length of supervision by medical experts can classify the failed examinees in the first trial. Cluster 1 corresponds to the group of examinees in which both the examinees and the medical experts supervising them have had a short career, Cluster 2 to examinees whose supervisors have had a relatively long career and Cluster 3 to examinees who have had a long career to some extent.

Since the classification by the K-means algorithm depends on the source data obtained by investigation, this classification might not be universal for all users of any insulin injection device. However, at least the data obtained in this investigation have the structure shown in Fig. 11.

	Total number of examinees	A: Relative Frequency (%)
Cluster1	215	36.9
Cluster2	265	45.5
Cluster3	103	17.7

Table 3. The total number of examinees in each cluster and its relative frequency.

1st research	The number of failed examinees	B: Relative Frequency (%)	Specialization Coefficient (B/A)
Cluster1	82	41.2	1.12
Cluster2	78	39.2	0.86
Cluster3	39	19.6	1.11

Table 4. The number of failed examinees in each clusters and its specialization coefficient (1st research).

2nd research	The number of failed examinees	C: Relative Frequency (%)	Specialization Coefficient (C/A)
Cluster1	117	44.3	1.20
Cluster2	103	39.0	0.86
Cluster3	44	16.7	0.94

Table 5. The number of failed examinees in each clusters and its specialization coefficient (2nd research).

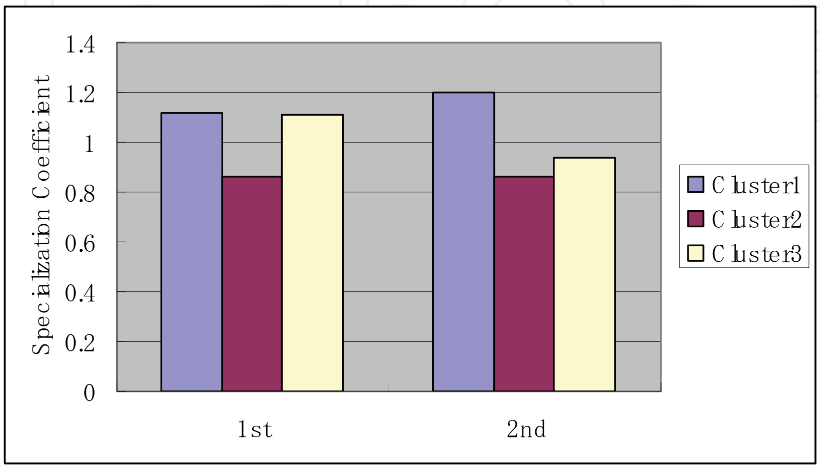


Fig. 12. The distribution of specialization coefficient in each research.

The specialization coefficients of the number of failed examinees are shown in Table 3, 4, 5 and Fig. 12. From these, we can see that Cluster 2 has less value than the other clusters. This can be interpreted as medical experts with a long career tending to succeed in guiding the examinees to accomplish the operation, and shows the importance of proper guidance by experienced experts.

It can also be seen that although in the first trial, Cluster 1 has as great a value as Cluster 3, in the second trial, Cluster 1 has a greater value than Cluster 3. This result suggests that, in the first trial, the career of the medical experts mainly has an effect on the result since it was the first time for examinees to use the device. In addition, in the second trial, the result can be considered as being affected by some examinees with a short use period of insulin injection device forgetting or omitting the operation since they do not recognize its importance.

The fact to be noticed is that the number of failed examinees in the second trial increased compared with the number in the first trial. This indicates that there is a general tendency to omit or forget the operation after a while, even if they first do it in the right way. It is important to notify patients of the correct method of using the device repeatedly.

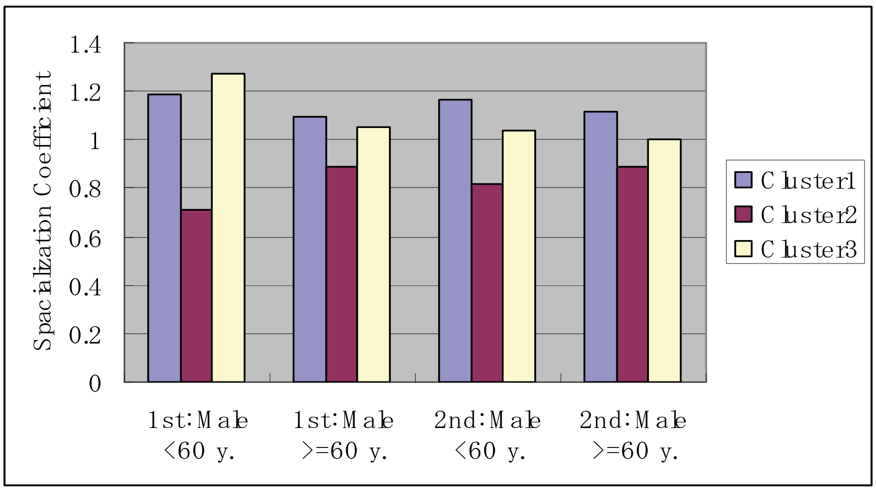


Fig. 13. The distribution of specialization coefficient for male examinees in each research.

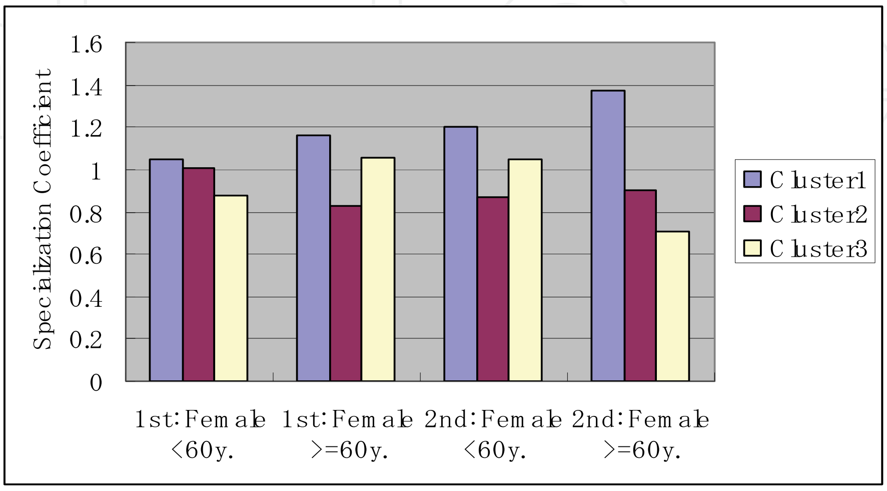


Fig. 14. The distribution of specialization coefficient for female examinees in each research.

Figure 13 and Fig. 14 show the results of analysis by gender and age. We divided the examinees into two groups by age, more or less than 60 years old, since the retirement of the principal or spouse usually changes the style of life and may cause a change in thinking and feeling about the operation.

There is a characteristic for male examinees younger than 60 years old in the first trial that Cluster 3 has the biggest value among the clusters. This shows that the rate of examinees with a long career who operate incorrectly is high. This suggests that the experienced examinees tend to omit the procedure due to the habituation of the pen-like injection device. Female examinees have the feature that the rate of examinees in Cluster 2 whose age is less than sixty in the first trial is higher than the other cases, and the rate in Cluster 1 of those over 60 years old is high in the second trial. This shows that, in the case of females younger than sixty, there is a tendency to fail for the first time even if the medical expert supervising them is fully experienced, and that, in the case of females older than sixty, they fail if they are not very used to the pen-type injection device and their supervisors have had a short career. This suggests that there is a tendency for young females to find usage of the new device not easy, and that older females tend to forget the correct operation after a while.

This indicates that, at the time of switching to a new device, young males need to be paid attention to if they have a long career and tend to operate by habituation, and young females should be paid attention to if they have a problem with a new device even if medical expert supervisor is fully experienced, and that older females need to be trained in the correct operation after a certain period time.

4. Summary and conclusion

In this chapter, we introduced some clustering algorithms, their variation to analyze questionnaires and investigations and their applications.

We first pointed out the weakness of agglomerated hierarchical clustering algorithm in application to multi-choice question, introduced selection co-occurrence measure to estimate the difference of how respondents select options and compared its result with the results of TwoStep algorithm method. The selection co-occurrence measure is applied to the vectors that correspond to the options and whose elements are assigned a value 1 or 0, depending on the response of the corresponding respondent. It is a distance-like index counting up the NAND value of each element of the two vectors. TwoStep algorithm supplied by SPSS is applied to the vectors each of which correspond to the respondents and whose elements correspond to each option and is suitable in the case to classify a number of respondents.

Next, we show the method to fix the suitable number of clusters in K-means algorithm, whose derivation has been done by trial and error in practice. We reformulated K-means algorithm as likelihood functions and applied Bayesian information criterion (BIC).

As the application of agglomerated hierarchical clustering algorithm with selection co-occurrence measure, we reviewed the analysis of the multiple-choice question part of the questionnaire about systemic transdermal absorbent preparations and identified the combinations of the simultaneously selected options that occur frequently as the set of opinions expressed by medical experts and obtained the results as follows:

- The combination of systemic transdermal absorbent preparations that medical experts deal with is mainly cardiac drugs and asthma drugs.
- The reason that medical experts select transdermal patches as a dosage form is that no damage is imposed on the gastrointestinal tract and that it maintains its effect for many

hours. In particular, we can make a good guess that the statement 'no damage is imposed on the gastrointestinal tract' comes from the fact that it is better for patients who have difficulty in eating or who have to take many oral drugs if the gastrointestinal tract is not affected.

- Many medical experts think that the design of the transdermal patch is desirable from the viewpoint of the prevention of medical accidents and that the mark should be integrated for the same efficacy and be widely recognized. The answer regarding the design can be attributed to the fact that, in such a situation, nobody can find what medicine the patch is after it has been put on the patient.
- Although pharmacists and nurses tend to answer that therapeutic classification mark and product name label should be promoted, doctors tend to be satisfied with current situation. This suggests that the people who deal directly with patients and medicines feel the necessity of the mark and the label strongly.

As the application of K-means algorithm whose number of clusters is determined by BIC, we review the analysis of the profile data of the examinees of the investigation to calculate how many patients can handle the injection device for antidiabetic drugs in the right way after guidance, and to identify difficult procedures for users. We classified the type of examinees who could not accomplish the procedure of 'confirmation by pulling the cartridge system before setting the units of the test injection', which has the lowest success rate, and obtained knowledge on the patients to whom information should be provided with special attention.

First, we applied the K-means algorithm to the profile data such as age, length of use of the pre-improved device, length of use of a pen-type insulin injection device and length of supervision by medical experts and calculated BIC to find the number of clusters approximating the data well. Next, in order to analyze the tendency of the failed examinees, we compare the specialization coefficient in each cluster obtained as the ratio of the relative frequency of the number of failed examinees to the relative frequency of all examinees. The investigation consists of two trials that are performed at certain intervals for the same examinee. We compared the results in each trial and found that, in the first trial, the career of medical experts mainly has an effect on the result, but, in the second trial, some examinees with a short use period of the insulin injection device forget or omit the operation since they do not recognize its importance. We also found that medical experts with a long career tend to succeed in guiding the examinees to accomplish the operation. This shows the importance of proper guidance by experienced experts.

We divided each cluster into examinee groups by gender and age (more or less than 60 years old) and analyzed the groups in the same way. It was found that, at the time of switching to the new device, young males with a long career need to be paid attention to if they tend to operate by habituation, and young females should be paid attention to if they have a problem with a new device even if the medical expert supervisor is fully experienced, and that older females need to be trained in the correct operation after a certain period time.

Of cause, it is insufficient only to analyze the data with which we deal in the above review. Since data mining is an activity to utilize data and to suggest improvement of operation or service, it is important to feed back the results to appropriate people and/or organizations. We fed back our results to medical experts and pharmaceutical companies by reporting the results and presentation at an academic conference. It is not necessarily easy to measure effectiveness of our feedback, since the investigations introduced in this review are conducted on a too broad scale to be conducted again as verification in the short term. However, we suppose that it is important to measure to what extent the results contribute to the improvement of the operations, e.g. better compliance to safely treat the injection device for antidiabetic drugs in our review.

5. Acknowledgement

The author would like to thank Dr. F. Tsuchiya and Prof. Dr. M. Ohkura for suggestion of the studies that are reviewed in this chapter and fruitful discussion.
The author would also like to thank Dr. H. Furukawa, TOA EIYO LTD., and Sanofi-Aventis K.K. for the valuable investigation data.

6. Reference

Berry, M.; Linoff, G. (1997) *Data mining techniques: for marketing, sales and customer support*, John Wiley & Sons Inc.

Kimura, M.; Ohkura, M.; Tsuchiya, F. (2006) Application of data-mining techniques to questionnaires about safety of drug use, *Proceedings of IEA2006 Congress* (CDROM).

Kimura, M.; Furukawa, H.; Ohkura, M.; Tsuchiya, F. (2006) Study on the safety of the usage of antidiabetic drug injection devices, *Proceedings of IEA2006 Congress* (CDROM).

SPSS inc. (2003). Clementine 8.0 Algorithms Guide, Integral Solutions Limited, pp.53-59.

Shimodaira, H.; Itoh, S.; Kubokawa, T.; Takeuchi, K. (2004). *Information theoretic model selection and its confidence evaluation* (in Japanese), a book in the series of Frontier of Statistical Science, Iwanami-Shoten.

Zhang, T.; Ramakrishnan, R.; Livny, M.(1996). BIRCH: an efficient data clustering method for very large databases, *Proceedings of the 1996 ACM SIGMOD international conference on Management of data*, pp.103-114.

Data Mining and Knowledge Discovery in Real Life Applications
INTECH

Data Mining and Knowledge Discovery in Real Life Applications
Edited by Julio Ponce and Adem Karahoca

ISBN 978-3-902613-53-0
Hard cover, 436 pages
Publisher I-Tech Education and Publishing
Published online 01, January, 2009
Published in print edition January, 2009

InTech Europe
University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

InTech China
Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821