Design of Elderly Behaviour Analytics Model in the Healthcare Industry in Hong Kong

Paul K.Y. Siu, *K.L.* Choy[*], and *H.Y.* Lam

Department of Industrial and Systems Engineering, The Hong Kong Polytechnic University, Kowloon, Hong Kong

Abstract. Due to the advancement of living standard and medical technologies, the life expectancy of people is further extended which brings tremendous impact to the society in the near future. The ageing population not only increases the pressure to public healthcare services, but also brings urgent needs in long term healthcare resources allocation planning in the society. This paper presents an Elderly Behaviour Analytics Model (EBAM) to identify the hospital healthcare service preferences of elderly for the future planning of healthcare industry. By conducting an elderly-targeted survey, the collected data is analysed to understand the factors affecting the decision of elderly to acquire healthcare services in hospitals. The model applies the genetic algorithm-guided clustering-based association rule mining approach for the segmentation of hospital service preferences of the elderly, and, the identification of relationship between personal characteristics within each cluster. This research study contributes to the understanding the actual healthcare needs of elderly which allows the government and healthcare service providers to adjust or modify the elderly policies and service content.

1 Introduction

Having a large population and high social ageing rate in the subsequent decades, China is being expected to reform the country in a tremendous degree and adapt to the foreseeable social changes with at least twenty years. As a city of China with around 7.4 million population, Hong Kong will also experience a massive local social impact. The proportion of local citizens aged over 65 will increase from fifteen percent in 2014 to thirty-six percent in 2064. Currently, Hong Kong has the world's highest population life expectancy and is not fully capable to handle the upcoming challenges arose from population ageing [1]. In order to maintain high level of social stability under the huge changes, a range of comprehensive and long-term preparations in the healthcare industry are necessary.

The sharp increase of ageing population in the society can cause the considerable reduction of local gross domestic product. Elderly welfare resources, especially those for elderly healthcare, are also expected to be rose accordingly. According to Vaapio et al. [2], people aged 65 or above are considered as elderly, while elderly's chronic diseases are generally more complicated to be handled than normal diseases. The amount of resources applied to serve an elderly patient is more than a double of that for a normal patient. Although the public hospital clusters are considerably large in scale and contributing over 75% in the hospital sector, elderly patients contribute over half of the total patient admissions in public hospitals. This reveals the

elderly relies much more on the public hospital healthcare services and resources than the people from other age groups. It denotes that the overall public hospital services demand, particularly the demand from elderly, are extremely high. To relief the burden of public hospitals, the investigation of elderly preferences on choosing hospital healthcare service is necessary to provide references and directions for the future development of healthcare industry in Hong Kong [3]. Understanding the actual healthcare needs of elderly allows the government and healthcare service providers to adjust or modify the elderly policies and service content etc. It is believed that the healthcare resources utilization can be enhanced through better resources allocation and also improve the service capability of hospitals and even the entire healthcare industry. Therefore, in this paper, an Elderly Behaviour Analytics Model (EBAM) is designed to identify the preferences among different types of elderly and facilitate comprehensive understanding on the elderly healthcare needs.

The remainder of this paper is organized as follows. Section 2 reviews the past literature regarding healthcare services in Hong Kong, elderly preference in healthcare services and data mining for behaviour analysis. Section 3 presents the design of the EBAM. Section 4 illustrates the use EBAM in Hong Kong. Section 5 discusses the results and implications of adopting the EBAM. Section 6 gives the conclusions.

---

[*] Corresponding author: kl.choy@polyu.edu.hk

2 Literature Review

2.1 Overview of Healthcare Services in Hong Kong

The healthcare industry in Hong Kong is well developed to provide professional medical services to its citizens and gives contributions to the high living standard and life expectancy of Hong Kong [4]. To reduce the severity of overwhelming in public hospitals, Hong Kong government has promoted a range of measures to relief the heavy burdens of public hospitals while some of the measures are found to be controversial [5]. For instance, increasing the fee charged for using accident and emergency services to reduce the situation of service abuse; renting inpatient beds in private hospitals to transfer patients from public to private hospital. In addition, some elderly-targeted measures are promoted to reduce the service usage by this major service user group. For example, the establishments of elderly health care voucher scheme and elderly vaccination subsidy scheme encourage elderly to regularly check their health and gain vaccine protection in clinics respectively [6]. In general, the local hospital sector, as the dominant stakeholder of elderly healthcare industry, is facing issues of resources shortage and misallocation. It requires detail evaluation for preparing to confront the challenges upcoming.

2.2 Elderly Preference in Healthcare Services

In order to meet the existing and future needs of elderly and their families, the investigation on the preference of elderly is needed [7][8]. Victoor et al. [9] discussed the situation in north-western European health-care systems and patients are encouraged to patients to make an active choice of health-care provider based on their preference. According to Al-Doghaither et al. [10], the data of sociodemographic variables (e.g. age, year of work experience) and hospital choosing attributes were collected for identifying factors that influenced the hospital selection. Discriminant analysis had been applied for identifying specific groups of patients. It was a common practice for studying the sociodemographic variables along with the significant influencing factors. It allowed clear discrimination of what choice a person is likely to make by studying the person's characteristics. Isroliwala et al. [11] investigated the perceived importance of hospital choosing attributes for both hospitals' general practitioners and patients in UK. The research showed significant differences between the two data sets. The hospital choosing attribute 'reputation or expertise of surgeon/consultant' was the most important hospital choosing attributes ranked by patients and it ranked the fifth important by general practitioners. This reflected that people are difficult to understand the needs and preferences of patients, even if the person is familiar with patients. Eliacin et al. [12] summarized that patient–provider relationship, fear of being judged, perceived inadequacy, and a history of substance abuse are the critical factors that influence patient preferences and involvement in treatment decisions.

2.3 Data Mining for Behaviour Analysis

K-means clustering is a data mining technique that separates data records into k clusters based on data similarity. The data similarity is generally determined by the Euclidean distance between data records and each cluster's mean value [13]. It is an effective and common practice for applying data clustering to reveal the structure of medical datasets. Khanmohammadi et al. [14] had clustered patients by their diagnosed diseases with the overlapping clustering. The results showed the simple grouping of the patient having one disease while the patient may also had suffered from other diseases. However, the choice of initial clusters' mean values would affect the clustering result. In order to minimise the effect of bias, assigning value in random basis is the ordinary approach to ensure the independence [15]. According to Paterlini & Krink [16], genetic algorithm (GA) applied in k-means clustering can significantly reduce the number of computation attempts and improve the optimization accuracy. It can generate the initial mean value of cluster in random basis to compensate the drawback of k-means clustering on highly depending on the initial mean values. It helps to achieve the optimal clustering result by using linear programming to assign data points into clusters according to their fitness function results. After that, the more favourable clustering arrangements would be kept for further evolution while crossover and mutation would be the processes to increase the diversity of clustering arrangement. The diversified data pool proceeds to a loop of procedures that calculate fitness function result, select the more favourable results and perform diversification, for evolving the clustering result.

Association rule is the technique for discovering association relationship among the data records in database [17]. The generated association rule provides meaningful association relations between elderly preference and their demand for supporting decision making in healthcare services [18]. Lee and Cheung [19] claimed that the co-occurrence of features in a given database can be discovered and converted into valuable information, using association rules. Apriori algorithm is the most classical and influential data mining algorithm in the association rules which is originally proposed by Agrawal et al. [20]. The algorithm can extract and generate k-itemset(s), which are frequently seen in database. Karabatak and Ince [21] designed an intelligent system with association rules and neural network to detect the pattern of breast cancer. Abdi and Giveki [22] proposed to diagnose erythemato-squamous diseases based on association rules, and with the help of support vector machine and particle swarm optimization. Yassine et al. [23] analysed the consumer behaviour by frequent pattern mining for health care applications. Wong et al. [24] developed an e-healthcare system using association rule mining to identify the association between the specific collected demographic data,

behaviour data and the health measurements data of the elderly.

To summarize, the elderly behaviour analytics model is needed for the segmentation of hospital service preferences of the elderly, and, the identification of relationship between personal characteristics within each cluster. Genetic algorithm-guided clustering-based association rule mining approach are proposed for analysing hospital service preferences. Clustering-based approach by k-means algorithm can separate preferences data into groups while GA can compensate the drawback of k-means clustering by evolving result in random basis. Meanwhile, association analysis is also integrated to the model as it can indicate the personal information correlation and characteristics.

3 Design of the Elderly Behaviour Analytics Model (EBAM)

In the model, all data acquired is formatted and stored in database. The corresponding set of data provided by each respondent is considered as a data record, also known as a respondent-representing data point. All stored data records are then imported to analysis tool for GA-based k-means clustering. The data cluster formation is determined by the distance between each data point and each cluster centre. The analysis terminates when the number of analysis execution reaches the maximum limitation to obtain the nearly optimal solution, while the value of each random chromosome and hence the cluster centres may be varied intentionally by crossover and mutation calculation algorithms. Once the clustering calculation has finished, the data points of each cluster are transferred to perform association analysis.

3.1 Data Collection and Storage Module (DCSM)

To initiate EBAM processes, it is necessary to acquire data and format it before analysis. The data of elderly healthcare needs is served as the primary input of the entire investigation and they are collected through questionnaires. Target respondents of the questionnaires are the elderly healthcare service users, specifically the Hong Kong citizens aged 45 or above. The questionnaires are distributed by both hardcopy and online version, with both Chinese and English versions, for catering various needs of respondents. The questionnaire consists of two major sections while each section responsible to collect data for clustering and association analysis exclusively. The first section composes of questions with Likert-type scales answer. This type of answer is suitable for clustering analysis application as it is a number within a scale and able to be quantified during analysis processes. The remaining section is mainly composed of questions with specific items answer, apart from those with Likert-type scales answer. Association analysis can cater all these different data type since it simply compares the item co-occurrence between data records. The processes of data extraction, transformation and loading begin after all

questionnaire responses are received. The translations of data coding and file format are implemented, in additional to the examination of response validation. The filtered and well-formatted data subsequently stores in centralized database.
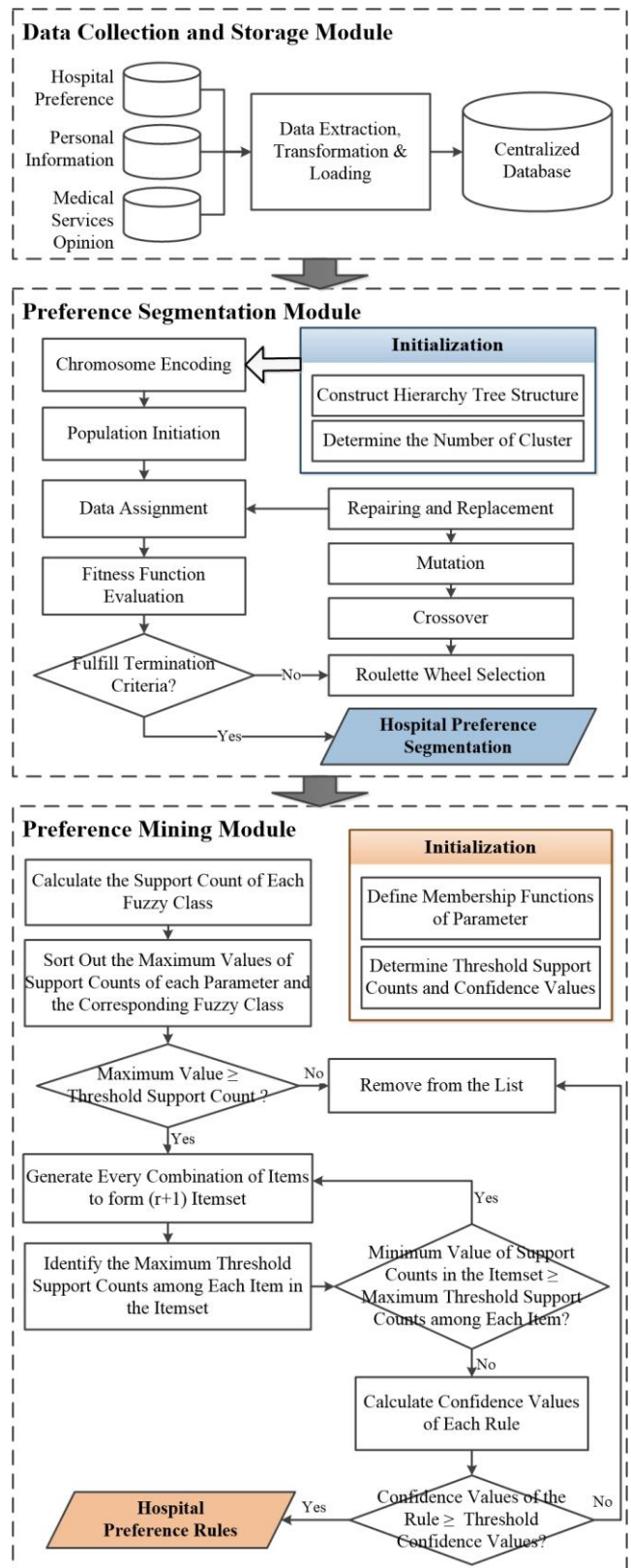

Fig. 1. Design of elderly behaviour analytics model.

3.2 Preference Segmentation Module (PSM)

GA-based k-means clustering is adopted for grouping data points with similar characteristics together by nearly optimal combination. The k-means clustering component responsible for data clusters assembly while GA component promoting result optimization. The Euclidean distance between each data point (data record) and cluster centre determines the allocation of the data point allocation among clusters. The data point would be assigned to the cluster which the Euclidean distance between them is the lowest among the distance between a data point and all cluster centres. Once the data and parameter preparation have completed, clustering analysis starts by random chromosome generation. Random chromosomes are variable arrays that represent the cluster centres. Fig. 2 shows the design of chromosome encoding for preference segmentation. Each chromosome consists of two regions, i.e. assignment region and parameter region. Value in the assignment region is either 1 or 0 to represent the selection status of the corresponding units in the actual data, while the value in parameter region contains positive integers to control the value of the corresponding unit in the actual data. Each unit in assignment region works in pairs with the corresponding unit in parameter region among the same chromosome, to serve as cluster centre. Moreover, the length of chromosome depends on the number of units of data string in database. The chromosome length will increase according to the number of cluster inputted in the previous step.

Fig. 2. Chromosome encoding for preference segmentation.

Fitness function is evaluated by computing the Euclidean distance between data points and cluster centres to facilitate clustering. The Euclidean distance between each data point and all the centres of clusters, thereby, is the calculated by the fitness function as shown in Eq. (1).

$$Dist_{i,j} = \left\| \left(\begin{bmatrix} Data_{i,1} \\ Data_{i,2} \\ \vdots \\ Data_{i,10} \end{bmatrix}^T - \begin{bmatrix} RandChromo_{s,10(k+j-1)+1} \\ RandChromo_{s,10(k+j-1)+2} \\ \vdots \\ RandChromo_{s,10(k+j-1)+10} \end{bmatrix}^T \right) \cdot \\ \times \begin{bmatrix} RandChromo_{s,10(j-1)+1} \\ RandChromo_{s,10(j-1)+2} \\ \vdots \\ RandChromo_{s,10(j-1)+10} \end{bmatrix}^T \right\|_2 \times Adjust_{s,j}. \quad (1)$$

Since the combination of more parameter values will bring significant increase in fitness value evaluation, an adjustment index is introduced in Eq. (2) for compensating the unfairness caused.

$$Adjust_{i,j} = \left(\sum_{s=1}^{p} RandChromo_{i,p(j-1)+s} \right)^{-1} \quad (2)$$

Fig. 3 shows an illustration of fitness function evaluation. The value of adjustment index is negatively correlated with the number of dimensions selected in indicator region of a random chromosome. The random chromosomes with fewer selected dimensions are given a larger adjustment index and one index is allocated to all chromosome per cluster.
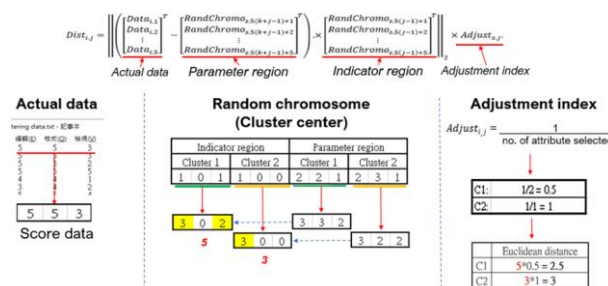

Fig. 3. Illustration of fitness function evaluation.

The Euclidean distance of all cluster centres is then calculated towards a same data point. The data points are assigned to the cluster with the smallest distance. After optimising the combinations in each cluster, the total distance value obtained by one chromosome is recorded as it represents one trial of data allocation in all clusters. It is recorded along with the optimal cluster-data point arrangement, as shown in Fig. 4.
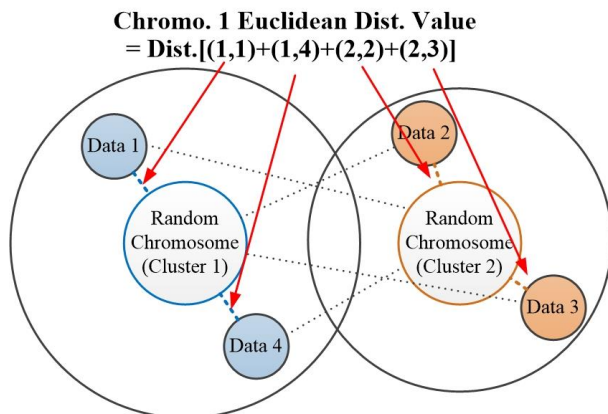
Fig. 4. Total Euclidean distance obtained by one chromosome.

The analysis will terminate eventually when the value of analysis execution counter fulfils the condition of maximum limitation set before. The optimal value of fitness function, the indicator and parameter region of the optimal random chromosome and the average computing duration per programme loop are displayed in the analysis tool. Data records of each cluster will be extracted according to the clustering analysis results for preference mining.

3.3 Preference Mining Module (PMM)

Association analysis is a data mining technique for discovering relationship between items by means of their co-occurrence frequency among database. It can identify association rules of data itemsets within a cluster established by GA-based k-means clustering. An association rule consists of two data components, condition and result. The percentage of a particular

condition and a particular result appear in the same data record can indicate the relationship stability between the two items. There are two percentage indicators that serve to verify the association rules: support and confidence. Support is the percentage of data records that include the specific condition among all records while confidence is the percentage of the condition-result combination occurrence among the records containing the specific condition. After determining the support and confidence values, Apriori algorithm is applied for identifying key association rules. The Apriori algorithm compares the support counts of every itemset in the data record list with support threshold, where support counts are formed by the total number of a particular item or itemset in the column. If the support counts of one item is lower than the support threshold, this item will be filtered and will not be able to proceed to the upcoming steps. After filtering the items separately, each remaining data is then combined with other items to form an itemset with 2 items. The support counts of each itemset is compared with support threshold for filtering the unqualified itemsets. The remaining 2 itemsets are combined to repeat the above comparison and filtration procedures until all columns of data records has completed the procedures or no itemset remains for proceeding to merge with items in next column. In case of all the record columns are examined, the support values of each condition are calculated. Hence, the confidence value of each candidate rule is obtained by dividing the support value of condition with that of candidate rule. Finally, the candidate rules, which their confidence values higher than the confidence threshold, are selected as the output.

4 Illustration of EBAM in Hong Kong

In this section, the EBAM is implemented for identifying elderly behaviours in the healthcare industry in Hong Kong. It includes data collection via questionnaire, GA-based k-means clustering for preference segmentation, and association analysis for preference mining.

4.1 Deployment of DCSM

Respondents are given a situation that they are currently having the needs of using hospital inpatient service. It is no doubt that the resulting factors from the overseas cases can only supply with limited efforts on filtering factors. Therefore, some local phenomena had been taken into consideration of factor selection, as for example: the overwhelming hospital wards environment and the storage of manpower in hospital. Meanwhile, some questionnaire drafts had been delivered to a small portion of target respondents for evaluating the effect of tentative factors. As result, 20 relatively influential factors are defined for studying the local hospital service selecting preferences of elderly. The hierarchical tree of hospital service preferences factors is shown in Fig. 5. The 20 factors cover 7 major dimensions of local hospital service, namely (i) service fee, (ii) service environment, (iii) service quality, (iv) hospital images, (v) staffing, (vi) hospital accessibility and (vii) financial aids coverage.

4.2 Deployment of PSM

MATLAB is used for programming the preference segmentation model. The flow of the programme is basically follows the calculation steps of k-means clustering while introducing GA for increasing the diversity of clustering results and hence obtaining near-optimal solution. This programme consists of three major procedures while each of them serves for random chromosome generation, core calculation and linear programming respectively. The flow of core calculation is shown in Fig. 6 and the parts for random chromosome generation and linear programming is highlighted in red and green respectively. Core calculation would first generate a pool of parent chromosome by the processes in the left region. Then, the processes in the right region form a loop and implement result evolution. The programme stops when the termination criteria is met.
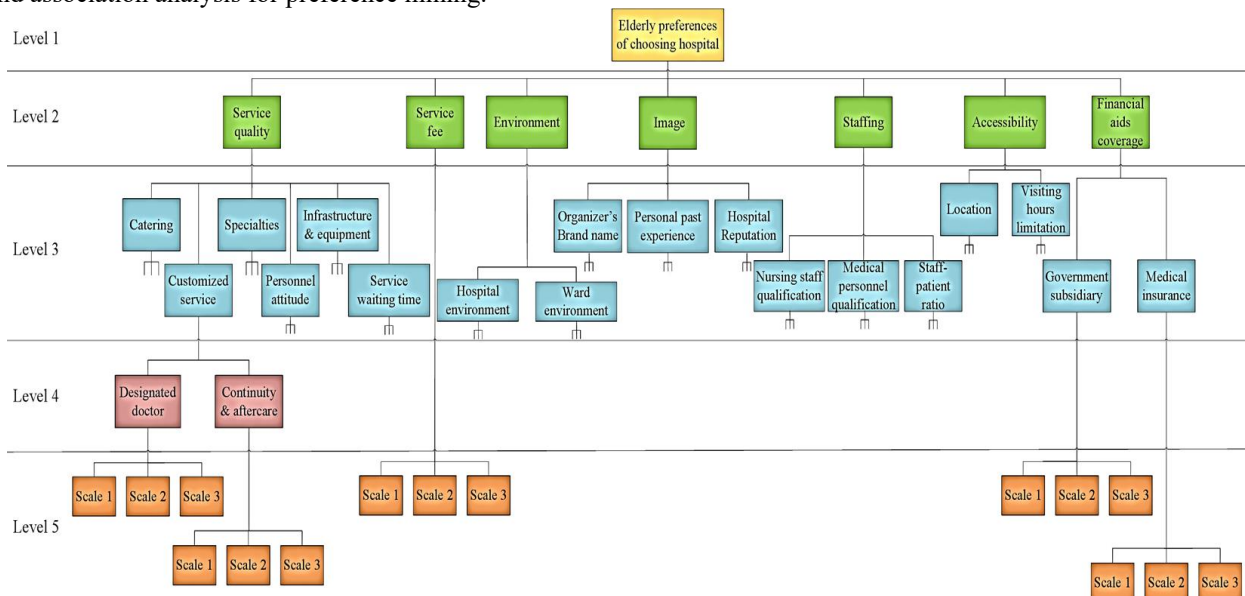

Fig. 5. Hierarchical tree of hospital preferences consideration factors.
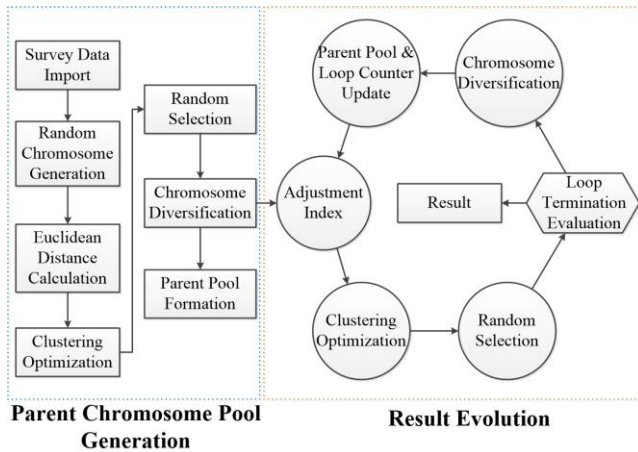
Fig. 6. Process flow of preference segmentation.

Once the processing of programme starts, it runs continuously and tries to achieve the most near-optimal solution. Termination of processing occurs when the execution of the programme loop reached the maximum execution condition. The obtained optimal value, the corresponding data-cluster allocation and chromosome are listed in the software as the GA-based k-means clustering analysis result, as shown in Fig. 7. The processing time fluctuates between every loop due to the random natures in chromosome formation, crossover and mutation.

4.3 Deployment of PMM

The association analysis depends on the data-cluster allocation result generated from clustering analysis. To identify significant association rules from numerous combinations in clusters, data mining computer software is necessary to be applied for analysis. Waikato Environment for Knowledge Analysis (Weka) is selected as the analysis tool that can greatly simplify the work by handling the calculation steps. Data records are extracted from database with reference of the data-cluster allocation in the clustering analysis result. These records are necessary to be transferred to the specially constructed file for complying with the standard of the association analysis tool to allow successful data importation. Attributes with abnormal responses need to be removed before conducting association analysis.

After all parameters are defined in the software, a list of association rules is generated. The minimum support and confidence values among all available rules are displayed on the top of the result. The shortlisted rules are listed according to the rule confidence values in descending order, as shown in Fig. 8. The values of improvement, leverage and conviction are also denoted along with the rules. The itemsets, which their support counts are less than support threshold, or the candidate rules which their confidence values are less than confidence threshold are eliminated automatically. All the generated rules must fulfil the support and confidence threshold constrains.
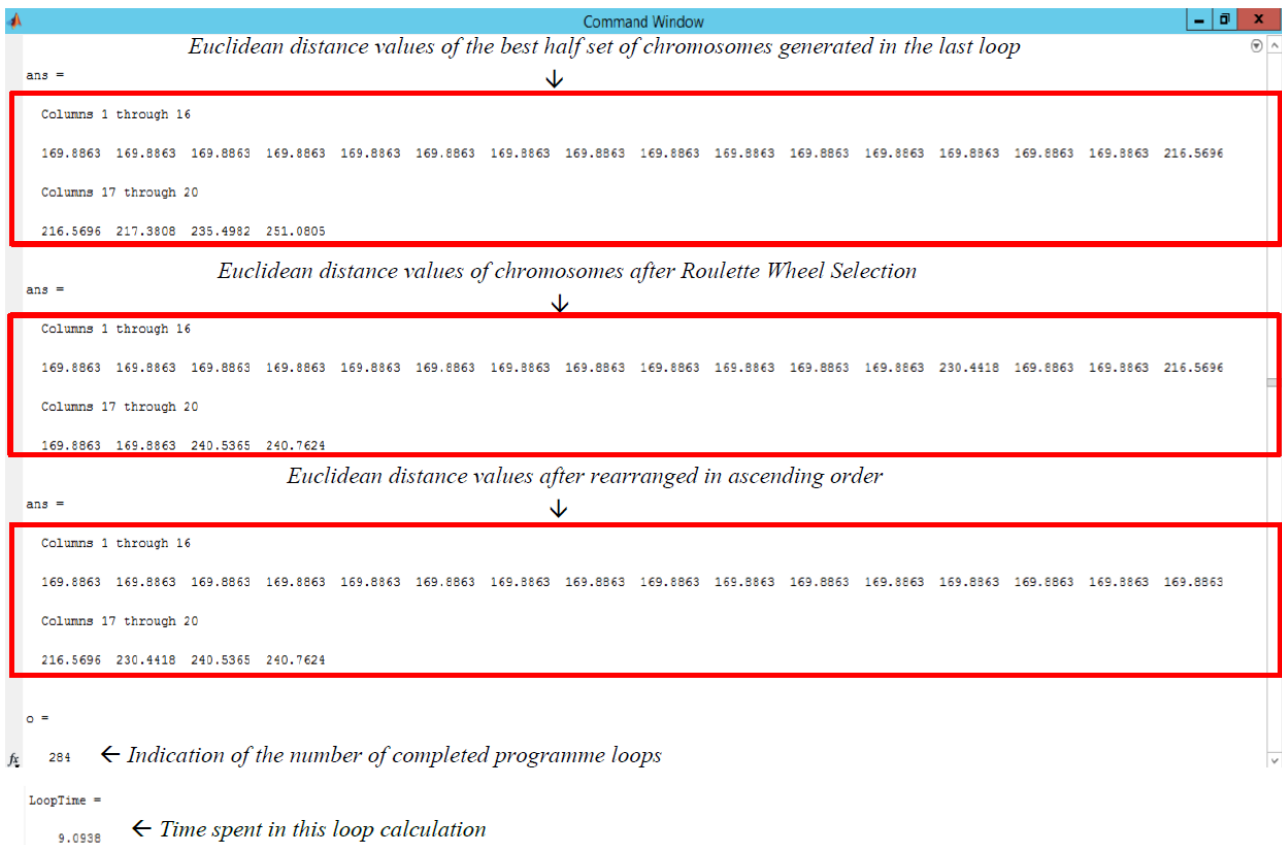

Fig. 7. Euclidean distance values of chromosomes after the completion of a programme loop.

```
=======

Minimum support: 0.35 (143 instances)        ←Overall performance of resulting
Minimum metric <confidence>: 0.8
Number of cycles performed: 13

Generated sets of large itemsets:

Size of set of large itemsets L(1): 12

Size of set of large itemsets L(2): 16          Resulting rules with corresponding confidence values on the left
                                                            ↓
Size of set of large itemsets L(3): 3

Best rules found:

  1. Willing_to_use_public_medical_service_over_private_medical_service=5 168 ==> Diagnosed_chronic_diseases=Chronic_disease_diagnosed 149    <conf:(0.89)> lift:(1.14)
  2. Number_of_current_medical_insurance=0 Willing_to_visit_clinic_over_hospital_for_medication_when_sufferring_from_mild_illness_e.g._flu=5 177 ==> Diagnosed_chronic_d
  3. Number_of_family_member_who_suffer_from_chronic_disease=0 Number_of_current_medical_insurance=0 178 ==> Diagnosed_chronic_diseases=Chronic_disease_diagnosed 154
  4. Diagnosed_chronic_diseases=Chronic_disease_diagnosed Number_of_body_check_per_year=0 178 ==> Number_of_current_medical_insurance=0 154    <conf:(0.87)> lift:(1.35)
  5. Number_of_current_medical_insurance=0 263 ==> Diagnosed_chronic_diseases=Chronic_disease_diagnosed 223    <conf:(0.85)> lift:(1.09) lev:(0.05) [18] conv:(1.43)
  6. Diagnosed_chronic_diseases=Chronic_disease_diagnosed Willing_to_visit_clinic_over_hospital_for_medication_when_sufferring_from_mild_illness_e.g._flu=5 186 ==> Numb
  7. Number_of_body_check_per_year=0 Number_of_current_medical_insurance=0 186 ==> Diagnosed_chronic_diseases=Chronic_disease_diagnosed 154    <conf:(0.83)> lift:(1.06)
  8. Willing_to_visit_clinic_over_hospital_for_medication_when_sufferring_from_mild_illness_e.g._flu=5 227 ==> Diagnosed_chronic_diseases=Chronic_disease_diagnosed 186
  9. Number_of_body_check_per_year=0 227 ==> Number_of_current_medical_insurance=0 186    <conf:(0.82)> lift:(1.27) lev:(0.1) [40] conv:(1.93)
 10. Diagnosed_chronic_diseases=Chronic_disease_diagnosed Number_of_family_member_who_suffer_from_chronic_disease=0 192 ==> Number_of_current_medical_insurance=0 154
```
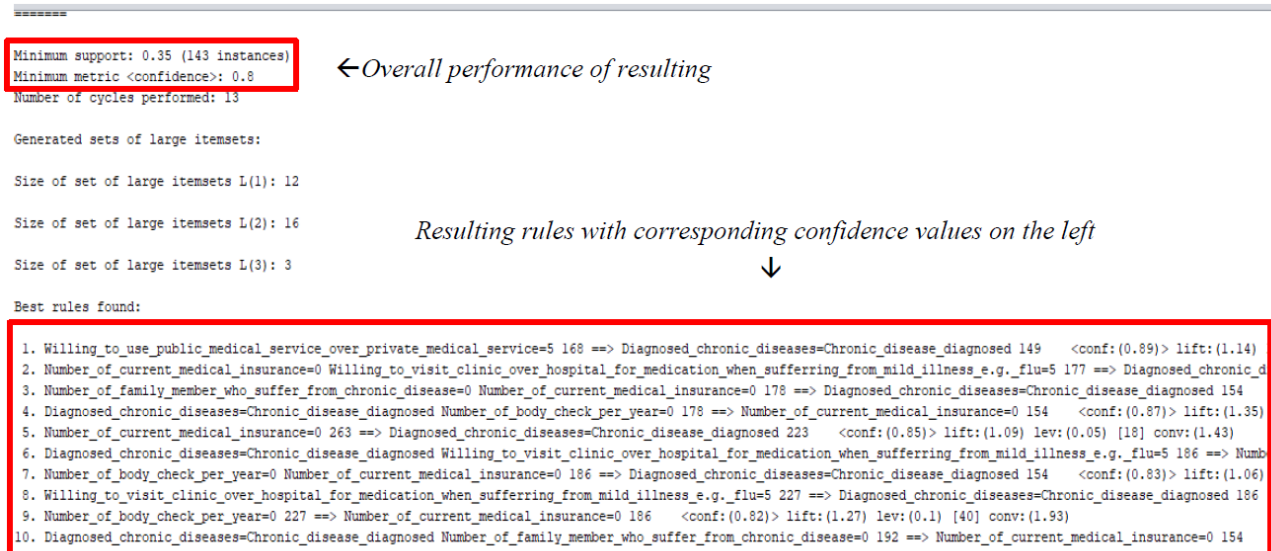
Fig. 8. Result of association rules generated.

5 Results and Discussion

5.1 Results of Elderly Behaviour Analysis

The clustering programme have been executed with a variety of variable inputs configurations, to achieve Euclidean distance values that are relatively close to the actual optimal value with intentional efforts in the random environment. The regular variable configurations of clustering analysis are 40 random parent chromosome, 2 clusters, 0.8 crossover rate, 0.1 mutation rate and 350 programme execution limit. In order to determine the best settings, different settings of variable configuration are used for analysis. Variable configuration and corresponding results of clustering analysis are shown in Table 1. The lowest Euclidean distance value 111.2394 is obtained in Trial 6 which has an additional cluster involved in data allocation. The increase of cluster generation number promotes the achievement of lower Euclidean distance value for better data record-classification. Thus, the similarity of the data records within their own clusters increases while the distances from the cluster centre, i.e. the mean value of the cluster, is shortened. The optimal clustering result is generated by dividing the data into three clusters. Table 2 shows the summary of analysis result.

The respondents assigned in Cluster 1 have a high level of requirement on selecting hospital. The mean value of 20 preferences attributes are all scored more than 3. That is, the respondents prefer all-rounded services coverage and highly concern on service quality. Generally, the differences between mean and mode values of attributes are low, which means most respondents share similar tendency on determining the impact caused by the factors. Yet the claimed impacts of 'Medical insurance' are not as consistent as in other attributes. Its mean value is 3.09 whereas the mode value is only 1. It implies that medical insurance is an important factor for a certain number of respondents.

Table 1. Variable configuration and corresponding results of clustering analysis.

Variable inputs	Optimal value obtained	Data-cluster allocation	Average processing per loop (second)
Regular setting (Test 1)	118.2137	(110,299)	10
Regular setting (Test 2)	120.0673	(306,103)	12
Crossover rate (0.8→0.2)	116.1349	(110,299)	6
Mutation rate (0.1→0.99)	169.8863	(368,41)	7
Number of parent random chromosome (40→20)	118.9488	(270,139)	3
Number of cluster (2→3)	111.2394	(268,78,63)	10
Maximum programme execution limit (350→250)	121.6935	(94,315)	8

Table 2. Optimal clustering result.

	Cluster 1	Cluster 2	Cluster 3
Attribute considered	Except "Medical insurance"	All	All
Number of data point	268	78	63
Characteristics (Mean value)	All attributes ≥ 3	Only direct-service related attributes (i.e. staff qualification, ward environment, equipment) ≥ 3	All attributes < 3

Respondents of Cluster 2 have clear and selective preferences for hospital selection. The mode values of all attributes are either in 1 or 5, while most of their values are close to their own mean values. The discrimination of assigning impact level to the factors is apparent. The respondents concern more on certain factors while not bothered by the others. The factors 'Hospital Specialties', 'Personnel attitude' and 'Hospital environment' have comparatively large differences between their mean and mode values. They indicate the

respondents have a wide range of claimed impact in these three attributes. To evaluate the result as a while, the respondents of Cluster 2 tend to concern about the factors that is directly related to the medical service. Hospital specialties, medical personnel and nursing staff qualification, staff-patient ratio, ward environment, infrastructure & equipment and continuity & aftercare are all factors critical to the service itself while they are being claimed to cause huge influence on the decision. For instance, ward environment has granted high influence level whereas hospital environment only has low impact as it is not important to inpatient service. Thereby, the preferences of Cluster 2 are relatively practical in terms of the service.

The respondents in Cluster 3 share a relatively low hospital selecting requirement. The mean values of all attributes are less than 3 and 19 out of 20 of the mode values are filled by 1. The hospital selecting decisions of respondents in Cluster 3 are extremely not easy to be affected by the factors covered in questionnaire. The most influential factor claimed by the majority respondents is 'Ward environment', with impact level 3 in mode value, while 'Infrastructure & equipment' has the highest mean value of 2.76. Apart from that, relatively large differences between the 2 values are observed in the factors 'Hospital specialties', 'Medical personnel qualification', 'Infrastructure & equipment' and 'Catering'. However, they do not give rise to huge effect to the extreme situation of low influence level for all factors.

The clustering result generated for three clusters is then imported to analysis tools for exploring the association rules inside a cluster. The support threshold is set to be 25% while the confidence threshold is set to be 85%. For cluster 1, none of the rules can fulfil the support and confidence threshold and hence no association rule is resulted. In cluster 1, the population size of data record in this cluster is relatively large (268). It requires a high level of data record similarity in order to fulfil the minimum support threshold. For cluster 2, the minimum support value of available rules is 0.9. It indicates the high similarity level within the cluster and the rules are valuable in representing the cluster characteristics. Eight rules are extracted as they can provide beneficial and valuable information for investigation. According to the generated rule, the three events, namely suffer from chronic disease, absolutely willing to visit clinic for mild illness treatment and absolutely willing to live in elderly home when necessary, are considered as interrelated. For cluster 3, the minimum support value of available rules is 0.65, which revels a medium level of similarities among data records in the same cluster. The minimum confidence value is above requirement and most of the resulting association rules are composed by 'Diagnosed chronic diseases = Chronic disease diagnosed', 'Number of body check per year = 0' and 'Number of current medical insurance = 0'. These items reflect that cluster 3 consists of data records similar to those of cluster 2, yet their meanings are different based on the different itemset combinations.

5.2 Discussion on the Preference Segmentation using GA-based Clustering Approach

By integrating GA into the clustering approach, it can tackle the limitations of traditional k-means clustering algorithm. Based on the literature review, it is found that the k-means clustering algorithm is highly sensitive to the initial cluster centre. If there is any isolated point in the analysis, the clustering result may be affected by calculating the average value of all objects. The use of GA can search for a nearly optimal solution for initial cluster centre, which could lower the impact of isolated points. In addition, with the use of GA-based clustering approach presented in this paper, the data are divided into clusters based on different set of attributes. That is, not all attributes are used as the cluster centre. GA has the ability to adjust the best combination of attributes for each cluster so that the choice of selected attributes can better represent the corresponding cluster. Fig. 9 shows the result generated using traditional k-means clustering algorithm. Without using the GA based approach, all attributes are used as the cluster centre. It is observed that the data are simply classified into three clusters with low, middle and high values of each attributes. Compared to the cluster result generated from GA-based clustering approach, no significant contribution is resulted for presenting the elderly preference with traditional k-means clustering algorithm. On the other hand, it is found that the cluster centre value of some attributes are quite similar. As shown in Fig. 9, the three cluster centre value of infrastructure and equipment, service waiting time, hospital environment, personal past experience and hospital reputation are similar. With similar values, it is difficult to decide whether it is suitable to include the data to the selected cluster.

```
Final cluster centroids:
                                             Cluster#
Attribute                        Full Data         0         1         2
                                   (409.0)   (101.0)   (139.0)   (169.0)
==========================================================================
Catering                           2.9756    1.2376    3.0576    3.9467
Designated_doctor                  3.4352    1.3267    3.2806    4.8225
Continuity_aftercare               3.9584    2.9802    3.6043    4.8343
Specialities                       3.1345    1.6634    2.8633    4.2367
PersonnelAttitude                  2.868     1.4752    2.6043    3.9172
Infrastructure_Equipment           4.2103    3.802     3.7554    4.8284
Service WaitingTime                4.1149    3.495     3.6906    4.8343
ServiceFee                         3.7628    2.2673    3.5324    4.8462
HosiptalEnvironment                4.0073    3.5743    3.5252    4.6627
WardEnvironment                    2.9046    1.396     2.9209    3.7929
OrganizerBrandName                 3.5941    2.2673    3.3885    4.5562
PersonalPastExperience             4.0269    3.3564    3.741     4.6627
HospitalReputation                 4.2054    3.7624    3.9065    4.716
NursingStaffQualification          3.1222    1.7525    3.1655    3.9053
MedicalPersonnelQualification      3.8631    2.7426    3.741     4.6331
StaffPatientRatio                  3.0562    1.4653    2.8777    4.1538
Location                           3.4866    2.703     3.4245    4.0059
VisitingHourLimitation             3.6357    1.6634    3.9065    4.5917
GovernmentSubsidiary               3.6088    1.9208    3.8201    4.4438
MedicalInsurance                   2.709     1.396     3.4892    2.8521
```

Fig. 9. Clustering Result without GA.

5.3 Discussion on the Analysis Results

5.3.1 Indication of Elderly Healthcare Needs

Investigating the healthcare needs of elderly is the major objective in this study. Three types of elderly needs can

be identified by integrating clustering and association rule results. Firstly, there is high demand of all-round and quality medical services for all elderly in Hong Kong. It is found that no significant personal characteristics could be identified based on their preference mining. Secondly, high quality and lean services are favourable for the elderly with chronic disease. They are willing to visit the clinic and live in elderly home when it is necessary since receiving high quality of medical services is their major concern. Thirdly, fundamental medical services with minimum resource is ideal for the elderly who do not have personal medical insurance and do not take regular body check. It is believed that this kind of elderly may have low awareness of staying health.

5.3.2 Implication on Future Development of Elderly Healthcare Industry in Hong Kong

With reference to the elderly-healthcare-needs interpretation, customization of elderly medical service is one of the important direction of future healthcare development in Hong Kong. In addition to paying concern to encounter the soaring service demand, allocating resource to effectively and flexibly match with dynamic service demand is the major measures in a long run. It is believed that the resource utilization can be optimised by the centralised organisation and control. Furthermore, educating citizens on the importance of taking regular body check annually is critical since respondents with no checking body in a year has a high occurrence rate of suffering chronic diseases as shown in the results of cluster 2 and 3.

6 Conclusions

Ageing population brings tremendous impact to the modern world. It is an unprecedented situation and would only be more severe in the future, due to the advancement of living standard and medical technologies. In this study, an elderly behaviour analytics model was designed for identifying elderly preferences of hospital service selection and personal characteristics of the elderly having specific preferences. The service preferences and personal characteristics were revealed by GA based k-means clustering and association analysis respectively while acquired first hand data with questionnaire survey. The needs of elderly healthcare and suggested direction for future development of the industry were obtained subsequently by the support of analysis findings. This research study contributes to the understanding the actual healthcare needs of elderly which allows the government and healthcare service providers to adjust or modify the elderly policies and service content.

References

1. J. H. Fong, Growing old in Hong Kong: responding to an increasingly frail population. In *Ageing in Asia-Pacific* (pp. 112-124). Routledge. (2018).

2. S. Vaapio, J. Puustinen, M.J. Salminen, T. Vahlberg, M. Salonoja, A. Lyles, S.L. Kivelä, Int J Gerontol, **9**(1), 34-39. (2015).

3. A.Q.L. Wang, B.H.P. Ng, L.P.C. Cheung, R.P.H. Chin, Hong Kong J Occup Th, **30**, 6-13. (2017).

4. X. Kong, Y. Yang, J. Gao, J. Guan, Y. Liu, R. Wang, B. Xing, Y. Li, W. Ma, J Chin Med Assoc, **78**(10), 569-573. (2015).

5. E.L. Wong, E.K. Yeoh, P.Y. Chau, C.H. Yam, A.W. Cheung, H. Fung, Soc Sci Med, **147**, 261-269. (2015).

6. A.H.Y. Lai, Z. Kuang, C.H.K. Yam, S. Ayub, E.K. Yeoh, Health Soc Care Community, **26**(3), 374-382. (2018).

7. T. Sandsdalen, R. Hov, S. Høye, I. Rystedt, B. Wilde-Larsson, Palliat Med, **29**(5), 399-419. (2015).

8. A. Mühlbacher, F.R. Johnson, Appl Health Econ Health Policy, **14**(3), 253-266. (2016).

9. A. Victoor, D. Delnoij, R. Friele, J. Rademakers, Health Expect, **19**(3), 667-678. (2016).

10. A.H. Al-Doghaither, B.M. Abdelrhman, A.W. Saeed, M.E.M.A. Magzoub, J R Soc Promot Health, **123**(2), 105-109. (2003).

11. S. Isroliwala, C. Wainwright, K.A. Sehdev, *Local View of Factors Influencing Patient Choice.* United kingdom: Cranfield University. (2004).

12. J. Eliacin, M.P. Salyers, M. Kukla, M.S. Matthias, J Ment Health, **24**(1), 24-28. (2015).

13. J. Qi, Y. Yu, L. Wang, J. Liu, Y. Wang, Int J Distrib Sens N. **13**(8). 1-17 (2017).

14. S. Khanmohammadi, N. Adibeig, S. Shanehbandy, Expert Syst Appl, **67**, 12-18. (2017).

15. M.Y. Cheng, K.Y. Huang, *26th International Symposium on Automation and Robotics in Construction, ISARC 2009*, 520-526 (2009).

16. S. Paterlini, T. Krink, Comput Stat Data Anal, **50**(5), 1220 – 1247. (2006).

17. D. Tomar, S.A. Agarwal, S. International Journal of Bio-Science and Bio-Technology, **5**(5), 241-266. (2013).

18. S.G. Alonso, I. De La Torre-Díez, S. Hamrioui, M. López-Coronado, D.C. Barreno, L.M. Nozaleda, M. Franco, J Med Syst, **42**(9), 161. (2018).

19. S.D. Lee, D. Cheung, Maintenance of discovered association rules. In *Knowledge Discovery for Business Information Systems* (pp. 173-209). Springer, Boston, MA. (2002).

20. R. Agrawal, T. Imieliński, A. Swami, A. *SIGMOD '93 Proceedings of the 1993 ACM SIGMOD international conference on Management of data*, 207-216, (1993).

21. M. Karabatak, M.C. Ince, M. C. Expert Syst Appl, **36**(2), 3465-3469. (2009).

22. M.J. Abdi, D. Giveki, Eng Appl Artif Intell, **26**(1), 603-608. (2013).

23. A. Yassine, S. Singh, A. Alamri, IEEE Access, **5**, 13131-13141. (2017).

24. B. Wong, G.T.S. Ho, E. Tsui, Ind Manage Data Syst, **117**(7), 1426-1445. (2017).