Discovery of Strong Association Rules for Attributes from Data for Program of All-Inclusive Care for the Elderly (PACE)

Shen LU
Soft Challenge LLC
Little Rock, AR, USA
slu@softchallenge.net

Alfred SEARS* and Joseph RADICH**
Dr. Sears' Center for Health & Wellness
Royal Palm Beach, FL, USA
{alsears*, jradich**}@alsearsmd.com

Richard SEGALL
Arkansas State University at Jonesboro
Jonesboro, AR, USA
rsegall@astate.edu

Thomas HAHN
University of Arkansas at Little Rock
Little Rock, AR, USA
tfhahn@ualr.edu

ABSTRACT

The Program of All-Inclusive Care for the Elderly (PACE) (2013)[6] study aimed to find out if the program we designed for the 11 month treatment can efficiently help people lose weight, and even can keep tracking of weight loss and body fat by checking some of the parameters we measured during the 11 months. We worked on the potentially significant parameters for weight loss in 11 months, such as age, height, weight, body size and body fat. We used association rule mining and classification rule mining to discover which parameters are significant for weight loss and what are the associations between weight loss and those significant parameters. Experimental results showed that weight loss with support from 0.2 to 0.9 and confidence from 0.7 to 1.0 is related to body weight and the changes of chest size, arm size, waist size, thigh size and hip size. In future, we will discover the associations among body weight, body size, body fat, heart beat and blood pressure.

Keywords: decision tree, information gain, association rules, classification rules, data analysis.

1. BACKGROUND

The Program of All-Inclusive Care for the Elderly (PACE) ((2013)[6], (2011)[9]) provides long term services and supports to Medicaid and Medicare enrollees. The PACE study is an ongoing project aimed at increasing health and life span. Many attributes such as age, height, weight, chest size, arm size, waist size, hip size, thigh size, total body fat, thigh fat, triceps fat and suprailliac fat were measured because if meaningful correlations could be found, then these attributes could potentially serve as health markers, which could be used to objectively quantify the relative general and subjective term "health". Since excessive weight is a risk factor for all kinds of age-related degenerative diseases, the effect of chest size, arm size, waist size, hip size, thigh size, total body fat, triceps fat and suprailliac fat on weight were investigated.

Association rule mining and classification rule mining are two different data mining techniques. According to Tan et al. (2006) [6], an association rule is an implication expression of the form X -> Y, where X and Y are disjoint itemsets, i.e. X Y= , and "a rule-based classifier is a technique for classifying records using a collection of 'if …then…' rules."

Association rule mining discovers all the rules with pre-defined thresholds, which can be used to estimate the strength of the association among parameters, and classification rule mining has a pre-defined target. In PACE project, the pre-defined target is weight loss. We used both association rule mining and classification rule mining to find associations between the properties above in order to draw conclusions about their possible interactions.

2. RELATED WORK

Hilderman et al.(1998)[3] mined association rules from market basket data using characterized itemsets for retail store example in a location where the customer base ranging from young families to the elderly. Jaroszewicz and Simovici (2002)[4] pruned redundant association rules using maximum entropy principle. Vannozzi et al. (2007)[7] extracted information on motor ability of elderly from clinical data through data mining. Papamatthaiakis (2010)[5] used association rules that produced a recognition accuracy of nearly 100% for modeled everyday home activities of the elderly.

3. DEFINITIONS

The following definitions of data mining techniques of itemset, sequence, support count, minimum support, candidate, frequent pattern and are from [2].

Definition 1. An **itemset** is a non-empty set of items. We denote an itemset by (x_1, x_2, ..., x_m), where x_j is an item.

Definition 2. Sequence is an ordered list of itemsets.

We denote a sequence s by $<s_1\ s_2\ ...\ s_n>$, where s_j is an itemset. We also call s_j an element of the sequence.

A sequence $<a_1,\ a_2,...,a_n>$ is a subsequence of another sequence $<b_1,\ b_2,\ ...b_m>$ for n m if there exist integers $i_1 < i_2 < ... < i_n$, such that, $a_1 \subset b_{i1}$,$a_2 \subset b_{i2}$, ... , $a_n \subset b_{in}$. For example, the sequence <(3)(4 5)(8)> is a subsequence of <(7)(3 8)(9)(4 5 6)(8)>, since (3) ⊂ (3 8), (4 5) ⊂ (4 5 6) and (8) ⊂ (8). However, the sequence < (3) (5)> is not a subsequence of < (3 5)> (and vice versa).

For simplicity, we assume that no data-sequence has more than one transaction with the same transaction-time, and use the transaction-time as the transaction-identifier. We do not consider quantities of items in a transaction.

Definition 3. Support count (or simple support) for a sequence is defined as the fraction of total data-sequences that "contains" this sequence.

Definition 4. Minimum support is the threshold, which is used to find the frequent itemsets. At this point, the majority of minimum support for discovering sequence patterns falls into two categories: those that are defined by customers, and those that are generated by algorithms. In this paper, we give a new method for finding a proper support in order to improve the performance of the whole sequential patterns mining procedure.

Definition 5. Candidate – the number of candidate items and the number of frequent items are important factors which significantly affect the performance of the mining procedure. Some algorithms try to generate only some length of sequences. Thus, we could balance the tradeoff between times wasted in counting non-maximal sequence versus counting extensions of small candidate sequences which is discussed in [1].

Definition 6. Frequent pattern Let X, Y be the set of frequent sequential patterns. According to the Apriori heuristic, since X is contained by Y, support(X) >= support(Y). The frequent pattern generated by the larger support must be the subsequence of that with the smaller support.

Definition 7. Association Rule [2] is considered interesting if they satisfy both a minimum support threshold and a minimum confidence threshold in [0, 1].

Definition 8. Correlation [2] If itemsets X and Y are both frequent but rarely occur together (i.e., $sup(X \cup Y) < sup(X) \times sup(Y)$),then itemsets X and Y are negatively correlated, and the pattern $X \cup Y$ is a negatively correlated pattern.

If $sup(X \cup Y) \ll sup(X) \times sup(Y)$, then X and Y are strongly negatively correlated, and the pattern $X \cup Y$ is a strongly negatively correlated pattern.

This dataset lend itself especially well to the data mining technique of discovering strong association rules. Clustering would have been theoretically possible too but couldn't be performed due to severe time constraints. Classification is not possible because all the tuples are not labeled.

4. DATA PROCESSING

Data from data files are available. One is body weight and body size dataset which has 70 samples, including the corresponding age, height, weight, body fat and body size in continuous 11 months and the other is weight loss dataset which has 57 samples, including the corresponding age, height, weight, body fat, body size in the first month, the second month, the sixth month and the ninth month and the weight loss. At the beginning of the weight loss, body weights can be changed very much, but, weight loss can really happen every several months, which is also proved by this project. Therefore, for the weight loss dataset, we measure weight loss twice in the first month, once in the second month, the sixth month and the ninth month in the one year. The dataset contains missing values. Therefore, extensive data preprocessing was necessary. Missing values were replaced by averages. The relatively small sample size of only 70 subjects is the consequence of this dataset being the results of a still preliminary study.

The reason for the relatively high percentage of missing values is due to the fact that the subjects (patients) didn't keep all their appointments, which they had for quantifying the previously described potentially health relevant attributes, especially towards the end of this study. However, since the reliability of the patient is to a large extend outside the control of the investigators, there is not much that could be done about this shortcoming except for - maybe - only selecting subjects (patients) for future studies, who kept most of their appointments in previous studies. That way the percentage of missing values could be reduced significantly. Moreover, if funds were available to pay subjects for their participation in similar future studies, they would have a much higher motivation to keep their appointments.

Another challenge was the high dimensionality (number of attributes) because if there are too many dimensions, we are getting too many candidates for frequent items, which is requiring too much computational resources thus quickly pushing personal computers beyond their computational limits. This often causes the computer to malfunction and hence could jeopardize the results of the entire data mining project altogether. Due to the high number of dimension the computer reached its hardware bottleneck surprisingly quickly and thus causing this data mining project to take unexpectedly long time.

We were able to generate strong association rules with RapidMiner® but so far we couldn't find a way to identify trends because we couldn't find any way with RapidMiner® to distinguish between positive and negative correlations of attributes. However, with regards to health markers and risk factors the direction of association, i.e. whether attributes are positively or negatively correlated, is of major importance.

4.1. How to overcome the challenges posed by many missing values?

There are many missing values in all of the tables. Since all of the attributes are continuous numbers, the RapidMiner® Data Transformation -> Data Cleansing -> Replace Missing Values Module to impute missing values with the average was used. For FP-growth algorithm, the tree structure required to transform continuous-valued attributes into nominal attributes, which were then converted into binominal attributes.

The RapidMiner® -> Data Transformation -> Type Conversion -> Discretization -> Discretize by Binning Module was used to discretize continuous-valued attributes into 15 bins. As Attribute filter type, "Value_type", for value type, were chosen. "All" and "range name type" were set to "long". Subsequently, RapidMiner® -> Data Transformation -> Type Conversion -> Nominal by Binominal module was used to convert nominal into binominal data. As attribute filter type -> "all" was selected, and "transform binominal" was selected.

4.2 Combining attributes to generate datasets

1. The "age", "height", "weight", body size and body fat attributes were combined to make a weight data set for 11 months.
2. The attributes "age", "height", “weight” and “weight loss” were combined to generate a "weight and weight loss" dataset.
3. The attributes "age", "height", “body size” and “weight loss” were combined to generate a "body size and weight loss" dataset.
4. The attributes "age", "height", “body fat” and “weight loss” were combined to generate a "body fat and weight loss" dataset.
5. The attributes "age", "height", “weight”, “body size”, “body fat” and “weight loss” were combined to generate a "weight, body size, body fat, and weight loss" dataset.

Thus, 5 datasets were created to discover previously unknown association rules.

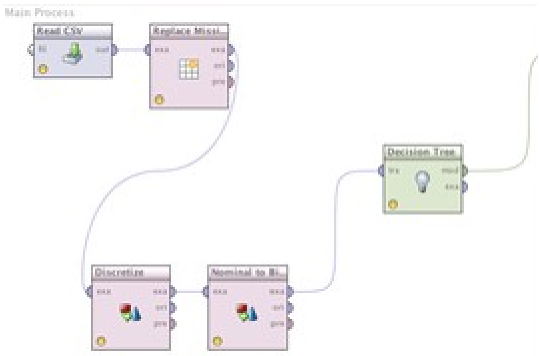
4.3 Creating RapidMiner® process by combining different modules

In order to find the significant parameters which can affect the body fat and weight loss, we built the hierarchy of the significant parameters by generating decision trees and discovered the associations among different parameters by generating association rules.

For decision tree generation, the whole process can be described as the following:

Step 1: Read CSV file
Step 2: Replace missing values
Step 3: Discretization
Step 4: Nominal to binominal
Step 5: Create a decision tree

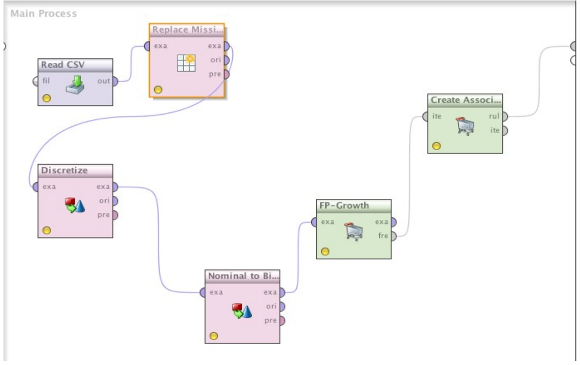
Figure 1 below is a screen shot of this 5-step process using RapidMiner®:



For association rules generation, the process can be described as the following:

Step 1: Read CSV file
Step 2: Replace missing values
Step 3: Discretization
Step 4: Nominal to binominal
Step 5: FP-Growth
Step 6: Create association rules

Figure 2 below is a screen shot of this 6-step process using RapidMiner®



5. DATA ANALYSIS

5.1 The hierarchy of significant parameters for weight loss

We divided weight loss attribute into 7 bins, as shown in table 1, such as [-∞, 5], [5, 10], [10, 15], [15, 20], [20, 25], [25, 30], [30, +∞]. In order to find the associations among weight loss, weight, body size and body fat, we generated four different datasets by combining different attributes together. The first dataset can be used to discover associations among weight loss and weights. The second dataset is about associations among weight loss, weight and body size. The third dataset is about associations among weight loss, weight and body fat. The fourth dataset is about associations among weight loss, body size and body fat.

Table 1. 7 classes for weight loss (grams)

Bins	min	max
Class 1	-∞	5
Class 2	5	10
Class 3	10	15
Class 4	15	20
Class 5	20	25
Class 6	25	30
Class 7	30	+∞

By combining weight loss, weights and body fat together, except weight in September, none of the attributes are significant and associated with weight loss. The same thing happened, when we combined weight loss, body size and body fat together.

By combining weight loss and weights together, the significant attributes are listed as WEIGHT-FEBRUARY, WEIGHT-INITIAL, and WEIGHT-SEPTEMBER.

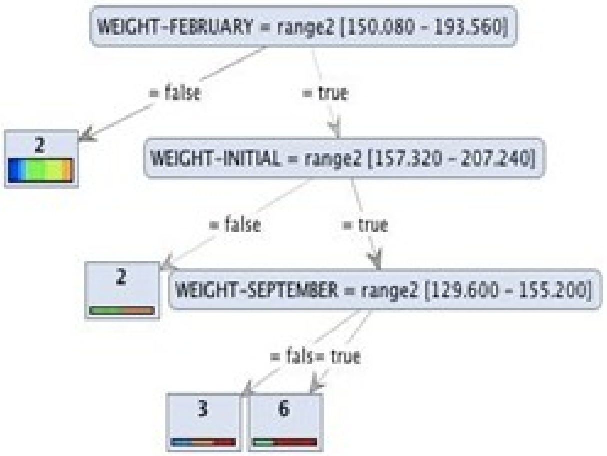

WEIGHT-FEBRUARY = range2 [150.080 - 193.560]
= false
= true
2
WEIGHT-INITIAL = range2 [157.320 - 207.240]
= false
= true
2
WEIGHT-SEPTEMBER = range2 [129.600 - 155.200]
= fals= true
3
6

Figure 3: The hierarchy of significant attributes about age, height, weights and weight loss.

Parameters	Initial	6 days	Feb	June	Sept
Weight	x		x		x
Weight loss	x	x	x	x	x

Table 2: Weight and Weight Loss Parameters

In the above Table 2, we shaded the boxes for the significant parameters which are the most associated with weight loss.

In Figure 3, we can see the associations among significant attributes, such as:

1. If the weight in February is not in range 2 [150, 193.560], the weight loss is in range 2 [5, 10]. That means if somebody is not very heavy, possibly, they will not lose a lot of weights.
2. If the weight in February is in range2 [150, 193.560] and the weight initial is not in range 2 [157.320, 207.240], the weight loss is in range 2 [5, 10]. Some people are not very heavy and in February if their weights are between 150 and 193.56, they will not lose a lot of weights.
3. If the weight in February is in range2 [150, 193.560], the weight initial is in range 2 [157.320, 207.240], and weight in September is in range 2 [129.6, 155.2], the weight loss is in range 6 [25, 30]. That means if somebody's weight starts between 157.32 and 207.24, in February between 150 and 193.56 and in September between 129.6 and 155.2, they will lose a lot of weights.
4. If the weight in February is in range2 [150, 193.560], the weight initial is in range 2 [157.320, 207.240], and weight in September is not in range 2 [129.6, 155.2], the weight loss is in range 3 [10, 15]. That means if somebody's weight starts between 157.32 and 207.24, in February between 150 and 193.56 and in September not between 129.6 and 155.2, they will not lose a lot of weights.

By combining weight loss and body size together, the significant attributes are ranked as chest in February, chest in September, waist in February, hip in February and arm initial.

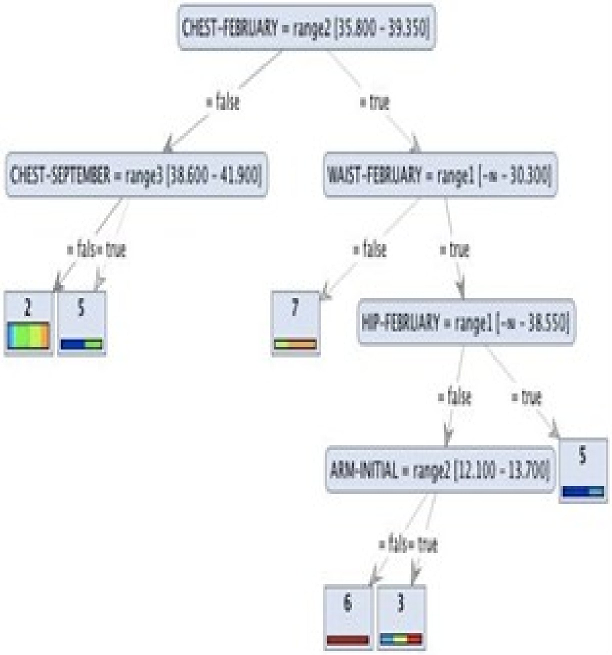

CHEST-FEBRUARY = range2 [35.800 - 39.350]
= false
= true
CHEST-SEPTEMBER = range3 [38.600 - 41.900]
WAIST-FEBRUARY = range1 [-∞ - 30.300]
= fals= true
2
5
= false
= true
7
HIP-FEBRUARY = range1 [-∞ - 38.550]
= false
= true
ARM-INITIAL = range2 [12.100 - 13.700]
5
= fals= true
6
3

Figure 4: The hierarchy of significant attributes about age, height, body size and weight loss

Table 3: Chest, Arm, Waist, Hip & Weight Loss

Parameters	Initial	6 days	Feb	June	Sep
Chest size			x		x
Arm size	x				
Waist size			x		
Hip size			x		
Weight loss	x	x	x	x	x

In the above Table 3, we shaded the boxes for the significant parameters which are the most associated with weight loss.

In Figure 4. we can see the associations among significant attributes, such as:

1. If chest size in February is not in range 2 [35.8, 39.350] and in September is in range 3 [38.6, 41.9], the weight loss is in range 5[20, 25].
2. If chest size in February is not in range 2 [35.8, 39.350] and in September is not in range 3 [38.6, 41.9], the weight loss is in range 2[5, 10].
3. If chest size in February is in range 2 [35.8, 39.350] and waist size in February is not in range 1 [-∞, 30.3], the weight loss is in range 7 [30, +∞].
4. If chest size in February is in range 2 [35.8, 39.350], waist size in February is in range 1 [-∞, 30.3], hip size in February is in range 1 [-∞, 38.550], the weight loss is in range 5 [20, 25].
5. If chest size in February is in range 2 [35.800, 39.350], waist size in February is in range 1 [-∞, 30.3], hip size in February is not in range 1 [-∞, 38.550], arm initial size is in range 2 [12.0, 13.000], the weight loss is in range 3 [10, 15].

6. If chest size in February is in range 2 [35.800, 39.350], waist size in February is in range 1 [-∞, 30.300], hip size in February is not in range 1 [-∞, 38.550], and arm initial size is not in range 2 [12.100, 13.700], the weight loss is in range 6 [25, 30].

5.2 **Looking for associations among the significant attributes**

Association rule mining can find all of the rules. However, we are only interested in the rules with significant parameters. On the decision trees, we can see significant parameters are weight-initial, weight-February, weight-September, chest-February, chest-September, waist-February, hip-February, arm-initial. All of association rules are listed in Table 2.

Except those association rules we find from the decision trees, according to support and confidence, the following rules of those significant attributes can also be found:

1. If weight initial is between 134.8 and 176.6, as shown in Table 3, with support 0.2 and confidence greater than 0.7, in month 2 weight is between 135.2 and 176.4, in month 3 weight is less than 157.7, and in month 4 weight is less than 153.2, in month2 hip size is between 35.1 and 40.95, in month 3 hip size is less than 37.8, in month 2 arm size is between 10.2 and 12.15.

Table 4: Associations about weight initial

weight initial [134.8 – 176.6]
weight month 2 [135.2 – 176.4]
weight month 3 [-∞ - 157.7]
weight month 4 [-∞ - 153.2]
hip size month 2 [35.1 – 40.95]
hip size month 3 [-∞ - 37.8]
arm size month 2 [10.2 – 12.15]

2. If weight in February is between 135.2 and 176.4, as shown in Table 4, with support 0.2 and confidence greater than 0.7, in month 1 weight is between 134.8 and 176.6, in month 3 weight is less than 157.7, in month2 hip size is between 35.1 and 40.95, in month 3 hip size is less than 37.8, in month 2 arm size is between 10.2 and 12.15.

Table 5: Associations about weight February

weight february [135.2 – 176.4]
weight initial [134.8 – 176.6]
weight month 3 [-∞ - 157.7]
hip size month 2 [35.1 – 40.95]
hip size month 3 [-∞ - 37.8]
arm size month 2 [10.2 – 12.15]

3. If chest size in month 2 is between 37.35 and 41.9, as shown in Table 5, with support 0.2 and confidence greater than 0.7, the chest size in month 3 is between 36.1 and 42.8, and the hip size in month 2 is between 35.1 and 40.95.

Table 6: Associations about chest February

chest-size-month 2 [37.35 – 41.9]
chest-size-month 3 [36.1 – 42.8]
hip-size-month 2 [35.1 – 40.95]

4. If waist size in month 2 is between 27.6 and 32.45, as shown in Table 6, with support 0.2 and confidence greater than 0.7, hip size in month 2 is between 35.1 and 40.95.

Table 7: Associations about waist February

waist-size-month 2 [27.6 – 32.45]
hip size month 2 [35.1 – 40.95]

5. If hip size in February is between 35.1 and 40.95, as shown in Table 7, with support 0.2 and confidence greater than 0.7, hip size in month 3 is less than 37.8, waist size in month 2 is between 27.6 and 32.45, weight initial is between 134.8 and 176.6, weight in month 2 is between 15.2 and 176.4, weight in month 3 is less than 157.5, chest size in month 2 is between 37.35 and 41.9, thigh size in month 2 is between 10.2 and 21.95, and arm size in month 2 is between 10.2 and 12.15.

Table 8: Associations about hip size February

hip size month 2 [35.1 – 40.95]
hip size month 3 [-∞ - 37.8]
waist-size-month 2 [27.6 – 32.45]
weight initial [134.8 – 176.6]
weight month 2 [135.2 – 176.4]
weight month 3 [-∞ - 157.7]
chest-size-month 2 [37.35 – 41.9]
thigh size month 2 [10.2 – 21.95]
arm size month 2 [10.2 – 12.15]

Table 9: Parameters & Months used in Experiments

Parameters\ Month	1	2	3	4	5	6	7	8	9	10	11
Weight	x	x	x	x							
Chest size		x	x								
Arm size		x									
Waist size		x									
Hip size		x	x								
Thigh size		x									
Weight loss	x	x	x	x							

According to those associations we discovered which included significant parameters, we found out that some parameters can also affect the weight loss, such as weight from the first month to the forth month, chest size in the second and the third month, arm size in the second month, waist size in the second month, hip size in the second month, thigh size in the second month, and thigh size in the second month. Right now, we only worked on parameters values in the first four months since we only have data for the first four months in PACE PB dataset. But, we can make sure those parameters can affect the weight loss all year long.

6. CONCLUSIONS

In this report, we presented the background of the PACE program, explained the reasons for looking at associations rules from PACE raw dataset, described the process of using PACE raw data to discover association rules and discussed the meaning of the rules. For classification rule mining, we used information gain to sort significant attributes in a hierarchical structure of attributes according to the significance of the attribute. For association rule mining, we set support 0.2 and confidence from 0.7 to 1.0 in order to explore how the attributes of age, height, weight, chest size, arm size, waist size, hip size, thigh size, total body fat, thigh fat, triceps fat, and suprailiac fat are associated with weight loss. Those rules we discovered in the project can be used to improve human's health. Future work, which could build on our findings, could include the discoveries of all the mentioned attributes especially with the attribute "body fat" since excessive body fat is increasing the risk of developing premature age-related degenerative diseases, such as diabetes.

7. REFERENCES

[1]. R. Agrawal and R. Srikant (1995). Mining sequential patterns. In ***Proc. 1995 Int. Conf. Data Engineering***, Taipei, Taiwan. March.

[2]. J. Han, M. Kamber (2000). **Data Mining: Concepts and Techniques**. Morgan Kaufmann Publishers. 2000.

[3]. R. J. Hildermann , C. L. Carter , H. J. Hamilton and N. Cercone (1998) "Mining Association Rules from Market Basket Data using Share Measures and Characterized Itemsets", ***Int. J. of AI tools***, vol. 7, no. 2, pp.189 -220.

[4]. S. Jaroszewicz and D.A. Simovici,(2002) Pruning redundant association rules using maximum entropy principle, ***Proc. Advances in Knowledge Discovery and Data Mining, 6th Pacific-Asia Conference (PAKDD'02)***, pp. 135-147, Taipei, Taiwan, May.

[5]. G. Papamatthaiakis (2010), Monitoring and modeling simple everyday activities of the elderly at home, ***Proc. Of 7th IEEE Consumer Communications and Networking Conference (CCNC),*** *Jan 9-12, pp. 1-5.*

[6]. P-N Tan, M. Steinbach, and V. Kumar (2006), **Introduction to Data Mining**, Addison Wesley, Boston, MA, pp. 207, 329.

[7]. G. Vannozzi, A. Cereatti, C. Mazza, F. Benvenute, and U. Della Croce (2007), Extraction of information on elder motor ability from clinical and biomechanical data through data mining, ***Comput Methods Programs Biomed***, Oct, v. 88, n. 1, pp. 85-94, Epub 2007, Aug 24.

[8]. The Program of All-Inclusive Care for the Elderly (PACE) (2013). http://www.medicaid.gov/Medicaid-CHIP-Program-Information/By-Topics/Long-Term-Services-and-Support/Integrating-Care/Program-of-All-Inclusive-Care-for-the-Elderly-PACE/Program-of-All-Inclusive-Care-for-the-Elderly-PACE.html. Centers for Medicare & Medicaid Services. 7500 Security Boulevard, Baltimore, MD 21244

[9]. Center for Medical Services (CMS) (2011), Department of Health & Human Resources (DHHS), **CMS Manual System, Pub 100-11 Programs of All-Inclusive Care for the Elderly (PACE)Manual**, http://www.cms.gov/Medicare/HealthPlans/pace/downloads/R1SO.pdf