Correlation Detection and Judgment of Abnormal Individual Behavior Based on Unsupervised Learning

Feng LI, Ting-ting LU and Ze-nan WANG*

1Room 125, Building No. 1, North Renming Road, Songjiang District, Shanghai, China
*Corresponding author

Keywords: Fuzzy clustering, Correlation detection, Anomaly judgment.

Abstract. In order to enable the Body Sensor Network (BSN) to detect the abnormal body state without the medical professional supervision, the method is proposed to find fuzzy association rules between daily exercise and physiological health data based on unsupervised learning. Firstly, the fuzzy clustering method is used to transform the numerical attributes to fuzzy categorical attributes. And then the frequent item set is calculated based on the support and confidence. The rules which reflect the behavior habit and laws are obtained from the final strong association rule through the minimum confidence screening. In order to avoid the useless rules and reduce the computational complexity, according to the characteristics of the exercise data, the frequent item set is divided into condition and conclusion groups. The abnormal state is detected by comparing the daily exercise data with the rules. The experiment with mass data had proven the validity of project.

Introduction

With the development of wearable equipment and related technology, many intelligent applications for motion recognition are introduced into the wearable device. C Mandery [1] proposed a dimensionality reduction of the feature space of the whole body motion recognition based on the Hidden Markov Model. In the eigenvector, considering higher dimension features such as joint angle, the algorithm proposed in [2] analyzed the time domain and frequency domain characteristics of the acceleration sensor data, and further improves the adaptability of the exercise state recognition algorithm by direction independence and step processing. And now there are many wearable devices are designed to collect the ECG and blood pressure signal too. Many methods on the processing of ECG signal in the human Body Sensor Network (BSN) are proposed in [3-5], such as the algorithm of heart attack based on Hidden Markov Model (HMM) and the method to remove the baseline drift of the ECG signal.

Based on the body physiological signal collected by the BSN, the individual daily exercise law, as well as the association relationship of these exercise with individual physiological signals, may be found. It can be very helpful to those who focus on daily exercise. The function, such as monitoring daily activities state, diagnosing of the abnormal state implemented on wearable devices, can highlight the value of the body sensor network (BSN). The key technique to complete the function is to find a kind of efficient and unsupervised learning algorithm. The BSN detected the changes of the individual body state, understand the relationship between the characteristics of the human body, and then complete the state monitoring and abnormal status diagnosis without the professional medical supervision. The exercise state and physiological data of a person are normal in most of the time, sudden changes often mean abnormal. Flowing the idea, the association rule mining algorithm is proposed to find the rules of individual daily exercise and physiological data based on the characteristics of individual exercise behavior in the paper. And then the method is discussed to discover the individual behavior or data anomaly based on the rules.

Against the market shopping basket, R. Agrawal used the association rules to express the knowledge in 1993, and proposed the Apriori algorithm to mine the association rules in the records.
in 1994, that is, to find hidden data patterns or knowledge from a large amount of accumulated data. At present the main researches on the algorithm are as follows: [7]

1) The multi-cycle mining algorithm. Its’ core idea is the "hierarchical algorithm"[8, 9]. Although the algorithm is simple and easy to understand, it may produce a great deal of candidate sets because of layer-by-layer iterative search, and require a large I/O load to scan transaction database multiply.

2) Dynamic counting algorithm. The advantage of the algorithms is the optimized performance with parallel I/O [10], but there are some problems in the application, such as sensitive to the data distribution and the choice of interval size. There is also error accumulation effect in the dynamic counting process.

3) Multi-valued attribute association rule mining algorithm [11], known as Boolean Association Rules. In the algorithm, the multi-valued attribute is divided into multiple intervals, each interval represents a property. However, the property does not represent the data distribution well, especially if the attribute value distribution is not uniform.

With the development of association rules, some scholars have found that Boolean Association Rules have some limitations in dealing with continuous variables. So the fuzzy mathematics is introduced into the association rules [12, 13]. Chan Man Kuok introduced the fuzzy rule into association rules to solve the problem of "hard borders"[13]. This is of great and practical significance to the development of association rules. K.Abhirami proposed a genetic algorithm [14] to find the rules that occur at the boundary of the fuzzy set, but it is more difficult to deal with problems with higher dimensions. Hangbin Li proposed a new tag-based model to solve the "rare project" problem [15]. Each item has a different label; the users can use the label combination to dig their own "rare items", which also lead to the high computational-complexity.

In the frequent item sets algorithm, in order to generate association rules, it is necessary to check all the subsets and then form a lot of useless rules too.

In this paper, in order to improve the efficiency of algorithm, according to the characteristics of mobile health data, a fuzzy clustering algorithm is proposed so that the attributes can be self-adjusted according to the change of data, and the number of rules, generated from the frequent item sets, is reduced by setting the limitation of condition and conclusions of association rules. This scheme is designed to further improve design quality and efficiency, not only gets the required rules, but also reduces the computational complexity.

Our research group have completed the wearable ECG equipment [16, 17], ECG automatic diagnosis [5] system and exercise state recognition [2] system. The study is carried out in the preceding work of our project group.

**Numerical Classification Based on Fuzzy Clustering**

The individual exercise data such as exercise status, exercise time, movement distance, number of exercise steps, and health data such as heart rate, respiratory rate and pulse rate are numeric data. Because the Apriori algorithm mainly deals with the discrete element aggregation, the numeric data should be converted to discrete element. Although the classification of the data according to its value is a simple solution, the traditional classification scheme of fixed category centers and distances is obviously unreasonable because the deviation is too large among individual’s data. The equal width division method is not ideal for the data with nonuniform distribution density, because the category attributes cannot reflect the semantics of the interval data without considering the relative distance between the data points or the interval. Moreover, it cannot meet the requirements of classifying in unsupervised learning.

Clustering analysis is an important unsupervised learning method. General clustering analysis methods include: clustering algorithm based on partition, clustering algorithm based on grid density and fuzzy clustering algorithm [18, 19]. The fuzzy C-Means (FCM) algorithm has been widely used because of its low computational complexity and fast convergence speed. FCM algorithm is sensitive to the initial center and lack of robustness. For the problem, many scholars have made improvements.
to the algorithm. In [20], a dynamic optimization FCM algorithm is proposed. In the algorithm the optimal classification number is automatically determined by the pseudo-F statistic (PFS) on the intra-class information and inter-class information. The dynamic clustering is achieved based on the sample density, and then is adopted in this paper.

The density parameter of the sample $x_j$ is defined as the follows:

$$\rho(x_j, R_j) = \sum_{i=1}^{N} u(R_j - d(x_i, x_j))$$

(1)

$$R_j = \frac{1}{c(N-1)} \sum_{i=1}^{N} d(x_i, x_j)$$

(2)

Where $R_j$ is the radius of the neighborhood of the sample $x_j$, $u$ is a function that calculate the degree of the sample $x_j$ within the radius of the sample $x_j$ neighborhood. $d(x_i, x_j)$ is the distance between $x_i$ and $x_j$.

FCM calculate the degree of the sample belong to a set, which determine the relation of the sample and an attribute. In this paper, a sample is a vector, the n vectors $(x_1, \ldots, x_n)$ is divided into c fuzzy sets. We should calculate the clustering center of each set so that the objective function reaches the optimal solution. The objective function is:

$$J_m(U, V) = \sum_{i=1}^{c} \sum_{j=1}^{n} (u_{ij}^m)(d_{ij}^2)$$

(3)

Where the $U_{ij}$ refers to the membership value, the degree of membership of $j$th element belongs to $i$th class, $d_{ij}^2$ refers to the distance between the $j$th element and the center point, the $J_m(U, V)$ represents the sum of the weighted distance of each point to each class. The $m$ is a parameter of the fuzzy degree. This algorithm has a constraint that the sum of degree of a sample membership value for all categories is equal to one.

A fitness function is designed based on the principle of minimum-in-cluster-distance and maximum-between-cluster-distance. The optimal solution is to make the objective function $J_m$ take the minimum value.

Let $D = \{I_1, I_2, \ldots, I_n\}$ be the characteristic attribute set which includes the user's exercise behavior feature. While the FCM algorithm is used to cluster each attribute, $D$ is transformed into a feature attribute classifier $D_f$:

$$D_f = \{F_1, F_2, \ldots, F_m\}$$

(4)

$$F_i = \{f_{i1}, f_{i2}, \ldots, f_{in}\}$$

(5)

Where $f_{ij}$ is a fuzzy set.

For example, the state vector set:

$D = \text{(state of exercise, duration of exercise, number of heartbeats per minute, distance of exercise, number of movement steps)}$

After fuzzy clustering, the following categories are formed:

Movement state $F_1 = \{\text{run, walk, still}\}$

Movement time $F_2 = \{\text{long, normal, short}\}$

Heartbeat $F_3 = \{\text{fast, normal}\}$

Movement distance $F_4 = \{\text{long, normal, short}\}$

Movement steps $F_5 = \{\text{more, normal, less}\}$
Individual Exercise Behavior Correlation Detection and Abnormal State Detection

Commonly there are two specified criteria for association rules: support and confidence.

Support and Confidence Calculation

(1) Support. The joint possibility of two events, for fuzzy sets the support, the attribute set $f_{ik}$ support for $f_{jl}$, is defined:

$$\text{Sup}(f_{ik} \rightarrow f_{jl}) = \left( \sum_{i=1}^{N} f_{ik}(d_s) \cdot f_{jl}(d_s) \right) / N$$

Where $d_s$ is the raw data in set $D$, $d_s \in D$, $N$ is the number of data in $D$. $f_{ik}(d_s)$ is the membership function of the data $d_s$ in the fuzzy set $f_{ik}$. For example, there are three samples in the set $D$ as shown in Table I, so the short-time $\rightarrow$ running's support is:

$$\text{Sup}(f_{i1} \rightarrow f_{23}) = \left( \sum_{i=1}^{3} f_{i1}(d_s) \cdot f_{23}(d_s) \right) / 3$$

$$= (0.8 \cdot 0.2 + 0.7 \cdot 0.7 + 0.2 \cdot 0.1) / 3 = 0.22$$

<table>
<thead>
<tr>
<th>Membership</th>
<th>Run(f11)</th>
<th>Walk(f12)</th>
<th>Long(f21)</th>
<th>Normal(f22)</th>
<th>Short(f23)</th>
</tr>
</thead>
<tbody>
<tr>
<td>d1</td>
<td>0.8</td>
<td>0.2</td>
<td>0.1</td>
<td>0.7</td>
<td>0.2</td>
</tr>
<tr>
<td>d2</td>
<td>0.7</td>
<td>0.3</td>
<td>0.2</td>
<td>0.1</td>
<td>0.7</td>
</tr>
<tr>
<td>d3</td>
<td>0.2</td>
<td>0.8</td>
<td>0.7</td>
<td>0.2</td>
<td>0.1</td>
</tr>
</tbody>
</table>

Table 1. Data in data set D.

(2) Confidence. The possibility of event $f_{jl}$ appears while event $f_{ik}$ appears in transaction $D$. In the fuzzy set, the confidence level is defined as follows:

$$\text{Con}(f_{ik} \rightarrow f_{jl}) = \frac{\text{Sup}(f_{ik} \rightarrow f_{jl})}{\text{Sup}(f_{ik})}$$

$$\text{Sup}(f_{ik}) = \sum_{i=1}^{N} f_{ik}(d_s) / N$$

The data in Table I, short-time-running confidence is:

$$\text{conf}(f_{i1} \rightarrow f_{23}) = \frac{0.8 \cdot 0.2 + 0.7 \cdot 0.7 + 0.2 \cdot 0.1}{0.8 + 0.7 + 0.2} = 0.40$$

Fuzzy Association Rules Mining

To expose Fuzzy association rules is to find relationship between the fuzzy concepts attributes. It is also to find the fuzzy frequent item sets in the attribute class library $D_f$ by the criteria: support and confidence. At present, Apriori algorithm [5, 6] or improved Apriori algorithm is the common algorithms of fuzzy frequent item sets. The main process is:

1) Convert the original database $D$ into the attribute category library $D_f$ by the FCM algorithm;

2) Get the fuzzy frequent sets from the data by the join and prune operation.

3) Get strong association rules from the fuzzy frequent sets

In the first round the fuzzy candidate sets are the items in the attribute class library $D_f$, while in the other rounds, the fuzzy candidate set are selected from fuzzy frequent set in the previous round by the join operation. And then obtain a frequent set from the fuzzy candidate sets by pruning operation. The pruning means that if the support of a record $T$ in the fuzzy candidate set is less than the defined minimum, it will be cut off. In addition, if a record $T$, one of its subset is not a fuzzy frequent set (support less than the threshold), will also be cut off. The loop’s termination condition is that the candidate set is no longer a fuzzy frequent set.
In the FCM, the associations’ rules are calculated from fuzzy frequent set through all the subsets. The time and space complexity of the process are very high. In order to reduce the time complexity, we divide the attributes in the fuzzy frequent set into two groups according to its actual meaning. One group only appears in the condition portion of the association rules, and the other group appears in the conclusion portion of the association rules. Such as the ‘number of heartbeats’ should only appear in the conclusion portion, and the ‘exercise time’ will only appear in the condition portion. The proposed method can reduce the computational complexity and the size of the association rules. Next we will pick out strong association rules based on confidence.

The fuzzy association rule mining algorithm is described as follows:

**Algorithm1:**

**Parameters:**

- \( F_1, \ldots, F_n \): the item sets in the database.
- \( f_{ik}(d_s) \): the membership function of the data \( d_s \) in the fuzzy set \( F_{ik} \).
- \( L_{k-1} \): the \((k-1)^{th}\) frequent item sets.
- \( C_k \): \(k^{th}\) candidate item set.
- \( C_{km} \): the \(m^{th}\) member of \( C_k \).
- \( F_{pre} = F_1 \cdot \ldots \cdot F_j \): the set of the condition of the rule.
- \( F_{con} = F_k \cdot \ldots \cdot F_n \): the set of the conclusion of the rule.
- \( \text{SubSet}(L_{k-1}) \): \((k-1)^{th}\) frequent subset of item sets.

**Input:** feature attribute fuzzy database \( D_f \), minimum support: \( \text{minsup} \), minimum confidence: \( \text{minconf} \).

**Output:** fuzzy association rule set.

**Begin**

\[
\text{If} \ (\text{Sup}(f_{ik}) \geq \text{minsup} ) \\
\text{The item set } f_{ik} \text{ is inserted into} \\
\text{the } k^{th} \text{ frequent item set } L_k \\
\text{For } (k=2; L_k \neq \emptyset; k++) \\
\text{Begin} \\
\text{Joint and generate the } k^{th} \text{ candidate item set } C_k \\
\text{Begin} \\
\text{For each } C_{km} \in C_k \\
\text{Begin} \\
\text{Sup}(C_{km}) = \sum_{i=1}^{n} \prod_{i=1}^{k} f_{ij} \\
\text{If} \ (\text{Sup}(C_{km} < \text{minsup}) ) \\
\text{delete } C_{km} \\
\text{End} \\
\text{The } k^{th} \text{ frequent term is } L_k = C_k \\
\text{End} \\
\text{For each } i \in \text{SubSet}(L_k) \\
\text{Begin} \\
\text{For } X \subseteq l \text{ And } Y \subseteq l \text{ And } X \subseteq F_{pre} \text{ And } Y \subseteq F_{con} \text{ And } (X \cap Y = \emptyset) \\
\text{Begin} \\
\text{If } \text{Conf}(X \Rightarrow Y) > = \text{minconf} \\
\text{Return } X \Rightarrow Y \text{ And Conf}(X \Rightarrow Y) \\
\text{End} \\
\text{End} \\
\text{End} \\
\text{End} \\
\text{End} \\
\text{End} \\
\text{End} \\
\text{End}
\]

**Finally,** the minimum confidence is used to screening the rules having potential value. The rules may be sorted by the confidence. And then we get the association rules of individual exercise behavior and physical data.

We designed the BSN system to collect the real-time in-exercise data, and then check if the data is consistent with current rules. If it is not consistent, that means abnormal state.
Experimental Analysis

In order to verify the rationality and validity of the algorithm mentioned above, in this chapter some experiments have been made to appraise the performance of algorithms. The parameters, the minimum support and the minimum confidence, were determined by experiments. The parameters of the experiment are shown in Table II.

<table>
<thead>
<tr>
<th>Parameters</th>
<th>Defaults</th>
</tr>
</thead>
<tbody>
<tr>
<td>number of attribute itemsets</td>
<td>5</td>
</tr>
<tr>
<td>Minsup</td>
<td>0.25</td>
</tr>
<tr>
<td>Minconf</td>
<td>0.96</td>
</tr>
<tr>
<td>items of the condition portion</td>
<td>Exercise State (p), Exercise Time (t), Motion Steps (s), Distance (d)</td>
</tr>
<tr>
<td>items of the Conclusion portion</td>
<td>Heartbeats (r)</td>
</tr>
</tbody>
</table>

The meaning of each attribute set in the experiment is shown in Table III:

<table>
<thead>
<tr>
<th>Attribute symbol</th>
<th>Attribute meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>p1</td>
<td>Walk</td>
</tr>
<tr>
<td>p2</td>
<td>Run</td>
</tr>
<tr>
<td>d1</td>
<td>long distance</td>
</tr>
<tr>
<td>d2</td>
<td>short distance</td>
</tr>
<tr>
<td>s1</td>
<td>more steps</td>
</tr>
<tr>
<td>s2</td>
<td>normal number of steps</td>
</tr>
<tr>
<td>r1</td>
<td>normal heart rate</td>
</tr>
<tr>
<td>r3</td>
<td>fast heart rate</td>
</tr>
<tr>
<td>t1</td>
<td>long time</td>
</tr>
<tr>
<td>t2</td>
<td>normal time</td>
</tr>
<tr>
<td>t3</td>
<td>short time</td>
</tr>
</tbody>
</table>

First of all, the database $D$ is transformed to fuzzy database $D_f$ by FCM clustered algorithm. Some results are shown in Figure 1.
Figure 1. Fuzzy clustering results.

Some membership functions of some heart rate data are shown in Figure 2.

Figure 2. Heart rate data membership.

After obtaining the fuzzy database and detecting fuzzy correlation, the first round of the candidate set are shown in Figure 3:

Figure 3. Candidate-1 itemsets.
In the first round, the candidate set contains the attributes \(\{d1, d2, p1, p2, r1, r3, s1, s2, t1, t2, t3\}\). By the joint and pruning operation the candidate items set is calculated in the second round (shown in Figure 4):

![Figure 4. Candidate 2 itemsets.](image)

In the second round the candidate set contains the attributes: \(\{d1p1, d1p2, d1r1, d1r3, d1s1, d1s2, d2p1, d2p2, d2r1, d2r3, d2s1, d2s2, p1r1, p1r3, p1s1, p1s2, p2r1, p2r3, p2s1, p2s2, r1s1, r1s2, r3s1, r3s2\}\). The candidate items in 3rd round contains \(\{d1p2r3, d1p2s1, d1r3s1, p2r3s1\}\), shown in Figure 5.

![Figure 5. Candidate-3 itemsets.](image)

We get the 3rd frequent item sets from the 3rd Candidate item sets. Next, the empty Candidate item set is got after self-join and pruning operation, so the algorithm terminates. The final fuzzy frequent set is \(S: \{'d1p2s1', 'd1r3s1', 'p2r3s1'\}\).

If the association rules are generated according to this fuzzy frequent set \(S\) directly, there are \(3 \times (C_1^4 + C_1^2)\) rules, most of which are meaningless. To avoid this problem, and reduce the computational complexity, the fuzzy frequent set is divided into the condition and conclusion groups. So the system will present user a simply association rules set. (Shown in Table IV)
Table 4. Comparison of generated exercise association rules.

<table>
<thead>
<tr>
<th>The previously generated association rules</th>
<th>Improved association rules generated after improvement</th>
</tr>
</thead>
<tbody>
<tr>
<td>d1-&gt;p2&amp;s1</td>
<td></td>
</tr>
<tr>
<td>p2-&gt;s1^d1</td>
<td></td>
</tr>
<tr>
<td>s1-&gt;p2^d1</td>
<td></td>
</tr>
<tr>
<td>d1^p2-&gt;s1</td>
<td></td>
</tr>
<tr>
<td>d1^s1-&gt;p2</td>
<td></td>
</tr>
<tr>
<td>p2^s1-&gt;d1</td>
<td></td>
</tr>
<tr>
<td>d1-&gt;s1^r3</td>
<td></td>
</tr>
<tr>
<td>s1-&gt;d1^r3</td>
<td></td>
</tr>
<tr>
<td>d1^r3-&gt;s1</td>
<td></td>
</tr>
<tr>
<td>d1^s1-&gt;r3</td>
<td></td>
</tr>
<tr>
<td>r3^s1-&gt;d1</td>
<td></td>
</tr>
<tr>
<td>p2-&gt;s1^r3</td>
<td></td>
</tr>
<tr>
<td>r3-&gt;p2^s1</td>
<td></td>
</tr>
<tr>
<td>s1-&gt;p2^r3</td>
<td></td>
</tr>
<tr>
<td>p2^r3-&gt;s1</td>
<td></td>
</tr>
<tr>
<td>r3^s1-&gt;p2</td>
<td></td>
</tr>
<tr>
<td>d1^s1-&gt;r3</td>
<td></td>
</tr>
<tr>
<td>p2^s1-&gt;r3</td>
<td></td>
</tr>
</tbody>
</table>

The confidence level of the improved exercise rules are shown in Table V:

Table 5. Confidence in the rules of exercise.

<table>
<thead>
<tr>
<th>Rules</th>
<th>Confidence</th>
</tr>
</thead>
<tbody>
<tr>
<td>d1^s1-&gt;r3</td>
<td>0.773370042885</td>
</tr>
<tr>
<td>p2^s1-&gt;r3</td>
<td>0.994355245943</td>
</tr>
</tbody>
</table>

The rule which confidence is less than the minimum threshold is deleted, and the finally strong association rule.

\[ p2^s1 \rightarrow r3 \]  \quad (11)

After the above steps, we can get some rules that reflect the law of a person's movement, such as continuous running in about 10 minutes, heartbeat is 80bpm. When the BSN find that today the people runs just 8 minutes, but the heartbeat is 100bpm, it will remind the user.

In order to verify the validity of the algorithm, about 60000 samples, covering 4 users (A, B, C, D) within 3 months, are collected. We've been testing the data, comparing the result, which the BSN system auto-check by the method above, with the data label. The measuring and analyzing results are shown in Table VI.

Precision P and recall R are defined as follow:

\[ P = \frac{TP}{TP + FP} \]  \quad (12)

\[ R = \frac{TP}{TP + FN} \]  \quad (13)

TP: the number of the samples which is judged abnormal while they are abnormal actually.
FP: the number of the samples which is judged abnormal while they are normal actually.
FN: the number of the samples which is judged normal while they are abnormal actually.

Table 6. Precision P and recall R.

<table>
<thead>
<tr>
<th>User</th>
<th>P</th>
<th>R</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>92</td>
<td>93</td>
</tr>
<tr>
<td>B</td>
<td>93</td>
<td>91</td>
</tr>
<tr>
<td>C</td>
<td>94</td>
<td>90</td>
</tr>
<tr>
<td>D</td>
<td>90</td>
<td>93</td>
</tr>
</tbody>
</table>
Conclusions
In order to enable the BSN to detect the abnormal body state of the user without the medical professional supervision, the method is proposed to find and analyze the relationship between exercise behavior and body state based on unsupervised learning. Firstly, the fuzzy clustering method is used to transform the numerical attributes to fuzzy categorical attributes. In order to reduce the computational complexity, according to the characteristics of the exercise data, before getting the association rules from the frequent set, the set is divided into condition and conclusion groups, and then obtain the final strong association rule through the minimum confidence screening.

The algorithm discussed in this paper focus on the individual state. In the future we are going to dig into the method to detect the abnormal state based on group data, not only the individual data. The study will focus on how to find the correlation between motion state and body data and then give suggestion on daily exercise.

Acknowledgement
Supported by National Natural Science Foundation of Shanghai: 16ZR1401100.

Reference


