A Healthy Aging Real-Time Monitoring System Based on Data Mining

Songtao Shang[1], Minyong Shi[2], Wenqian Shang[3]

[1, 2, 3] School of Computer Science, Communication University of China, Beijing, China

Email: songtao.shang@foxmail.com

Keywords: Data Mining; Naïve Bayes; K-Means; KNN; Population Aging

Abstract. Aging population is serious problems now in China. Many aging people have only one child, their children have many works to do every day. The only children have little time to care for their parents' health and daily life. This paper introduces a real-time monitoring system which can collect the vital information through wearable devices and can analyze the data in real-time. The children can monitor their parents' health in real-time by using mobile devices. Data mining technology plays a key role in the system. All the aging people's health data can be analyzed by using data mining technology. This paper uses Naïve Bayes, K-Means, KNN algorithm and so on to analyse the aging people's health data.

1 Introduction

In recent years, aging population is becoming a serious problem in China. At the same time, the generation of only one child has gradually entered the aging stage. Their children are busy with work, and have little time to care for their parents' life and health. According to the report of "Twenty-first Century: to celebrate the challenge but also the aging", by 2030, people aged 60 or above will be increased to 2 billion; in 2100, it will reach nearly 3 billion [1]. For the only children, how to care about their parents' daily life is becoming an important issue. Under this circumstance this paper introduces a system to help the only children to care about their parents' life and health. This system is divided into two parts: front part and back end. The front part is web page which can be logged in by a computer or an APP which is applied for Android or iOS. The back end involves a lot of technology of data mining, data collection and data processing.

The system is a comprehensive system and involves a lot of technology. The data collecting sub-system is responsible for collecting the vital information of aging people. The information display sub-system is to display a variety of policy information and aging people's activities which is using web crawler [2][3] to search from the Internet. The data processing sub-system is responsible for analyzing the massive data. This is the core of the system. The results are an indicator of the aging people health. The results are also support for the third party application to develop their new products. We also use Data Mining [4][5] technology to process the data. The other important sub-system is wearable devices [6][7], which can gather the vital information of aging people, such as blood pressure, pulse, amount of exercise, and so on.

By using the system, the aging people can browse the latest policy news or aging activities, download their interested mobile phone games from the app store, and express their views on the forum. The system provides a platform for the aging people to contact with each other. On the other hand, the children can monitor their parents' health in real-time through the system in the remote. So, their children have much time to do more work.

2 System Architecture

Figure 1 shows the system architecture. The System includes Data Collection, Data Base, Data Processing, Wearable Devices, and third party applications. The purpose of the system is to provide services for the only children and their parents. The function of the system is flexible and various. The wearable devices can gather the vital information of aging people, and the data is stored in the databases. The data will be analyzed and the results are stored into the database too. The aging people or the only children can log in the system to review the results. Otherwise, the results can be

pushed to the only children's mobile phone or iPad. It is very convenient for the only children to keep an eye on the status of the aging people's health.

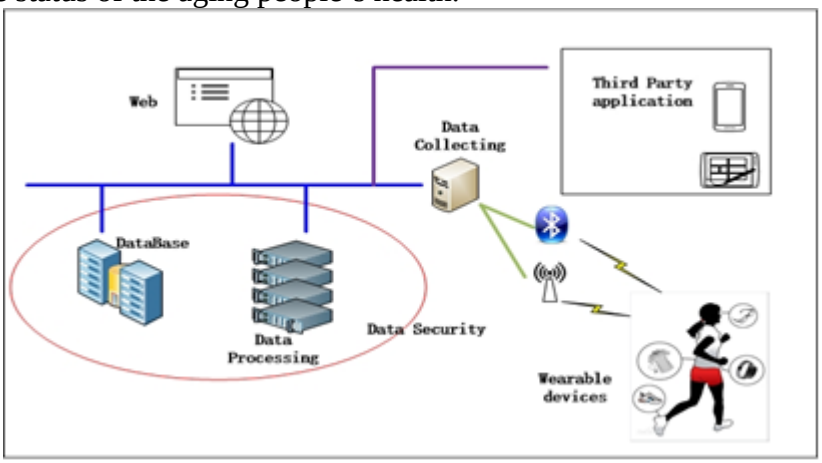

Figure 1 System Architecture

1. Database: All the collected data and analyzed data are stored into the database. We use SQL database and NoSQL database. The hot data and online data are stored in SQL database. The offline data and processed data are stored in NoSQL database.

2. Data Processing: When time passed, a lot of data are stored into the database. We use Big Data processing framework to analyze the data, the Hadoop framework is a good choice. It is a parallel computing model suitable for Big Data processing.

3. Wearable devices: Wearable devices are the electronic devices that can be worn on human body. The commonly devices can be shown as follows:

Google glass, it is like a glasses wearing on the human's head. It can record what the human has seen. This device can be developed again to expand its function.

Electronic bracelet or foot ring, it is worn on the human's wrist or ankle. It is designed to detect the human's blood pressure, pulse, monitoring sleep information, recording the steps number of movement and other information. All the data reflect the status of the health of human.

The wearable devices transfer their data to the central computer by WiFi or Bluetooth. So, we need to build some WiFi or Bluetooth stations in the community where the aging people lived in.

4. The third party application: This part is to support for the third party to develop new application for aging people and their children. For example, some companies can develop the application based on the processed data on Android or iOS platform. The application can push the aging's health results to their children's mobile devices. Their children can monitor their parent's health in time. The aging people also monitor themselves health status on their own phone.

5. Data security: The health data is individual privacy. The data security assures that each person can access to their own data.

6. Web pages: This part is to support aging people and their children to log in the system. From the web page we can see the aging's health data or status. The web page is also a platform. The aging people could browse the new policy news or express their views through the platform.

3 The Core Algorithm

The main idea of the system is to provide services for the aging people. According to the data that we collected and analyzed, we can find the probability of potential disease of the aging people. The main algorithm of the system is data mining algorithm. Through the analysis of historical data, we can mine the probability of aging people's diseases in the future, so as to improve the quality of the aging people. In this paper, we use K-means algorithm [8], Naïve Bayes algorithm [9], and KNN algorithm [10], and so on.

3.1 K-Means algorithm

K-Means algorithm is a simple and easy way to classify a given data set through a certain

number of clusters (assume k clusters). The main idea is to define k centroids, one for each cluster. The K-Means algorithm chooses the centroid randomly from the data set. Then, take each set belonging to a given set and associate it to the nearest centroid.

Given a set of observations (x1, x2, …, xn), where each observation is a d-dimensional real vector, K-Means clustering aims to partition the n observations into k (≤n) sets so as to minimize the within-cluster sum of squares. The K-Means clustering partitions a data set by minimizing a sum of squares cost functions.

$$D = \sum_{j=1}^{k} \sum_{i=1}^{x} \| x_i^{(j)} - c_j \|^2 \qquad (1)$$

Where $\| x_i^{(j)} - c_j \|^2$ is a chosen distance measure between a data $x_i^{(j)}$ and the cluster center c_j, is an indicator of the distance of the *n* data from their respective cluster centroids.

3.2 Naïve Bayes algorithm

Naïve Bayes assumes that the value of each feature has independent influence on a given class, and this assumption called lass condition independence that used to simplify the computation, and in this sense, we call it "Naïve". Naïve Bayes is based on Bayesian theorem.

We assume that d is a data sample with unknown class label and H' is an assumption. If data sample d belongs to a particular class c, for the problem of categorization, we hope to get $P(H'|d)$. Namely, we hope to know the probability of H' when data sample d is given. $P(H'|d)$ is a posteriori probability or a posteriori probability under the condition of d.

AS we know that, $P(d)$, $P(H')$ and $P(H'|d)$ can be calculated from the given data. The Bayesian theorem provides a method for calculated from the given data. So, Bayesian theorem can be described as follows:

$$P(H'|d) = \frac{P(d|H')}{P(d)} \qquad (2)$$

Each data sample is represented as an *n*-dimensional feature vector that describes *n* measures of *n* samples. Assumed *m* class of c_1, c_2, …, c_m and given an unknown data sample *d* (no class label), they will be sorted into the class which has the highest posteriori probability based on categorization. In other words, a native Bayesian classifier will assign unknown sample to the class c_i, if and only if : $P(c_i|d) > P(c_j|d), 1 \le i, j \le m, j \ne i$.

Thus, we can maximize the $P(c_i|d)$, where class c_i has the largest $P(c_i|d)$ and is called the maximum posteriori assumption. According to Bayesian theorem:

$$P(c_i|d) = \frac{P(d|c_i)}{P(d)} \qquad (3)$$

Since $P(d)$ is a constant for all classes, we only need to maximize $P(d|c_i)P(c_i)$. If the prior probability of the class is unknown, it is usually assumed that the probability of these class is equivalent, that is $P(c_1) = P(c_2) = ... = P(c_m)$. So we maximize $P(d|c_i)$ only. Otherwise, we should maximize the $P(d|c_i)P(c_i)$. We also know that the prior probability of a class can be calculated by $P(c_i) = \frac{s_i}{s}$, where s_i is the number of training samples of the class and *s* is the total number of training samples.

It may cost much to calculate $P(d|c_i)$ when the given data sets with many attributes. To reduce the computational cost of $P(d|c_i)$, we can simply assume that the class is conditional independent. If we know the class label of a sample, and assume that the value of each property is conditional independent, namely, there is no dependent relationship between every pair of properties. Hence:

$$P(d|c_i) = \prod_{j=1}^{n} P(x_j|c_i) \qquad (4)$$

3.3 The improved Naïve Bayes

We know that feature weight is another important aspect for classifier. So in this paper, we

combine the feature weight with Naïve Bayes classifier, to improve the precision of the classifier.

Feature weighting has the following three general steps:

1. Calculating the ability of distinguish for each feature;
2. Screening a certain number of features according to the ability distinguish;
3. Adjusting the weights of features, emphasizing the features with a strong ability to distinguish, and inhibiting the lower or no one.

There are two ways to execute Step 2:

Method 1, setting a threshold of assessment and deleting the features below the threshold.

Method 2, setting a threshold of retained number of features, sorting the features by assessment and retaining the top predetermined number of features.

Step 3 is to construct a strategy to adjust the weight. Weight adjustment aims to highlight important features and inhibit the secondary ones.

TF-IDF is a commonly used function of feature weight adjustment, but the IDF function in the TF-IDF cannot reflect the feature's importance well. Therefore, we use a feature evaluation function to replace the IDF function and construct a new feature weight function, TF-TWF function. TWF represents a feature evaluation function, the TF-TWF weighting formula is as follows:

$$W_t = TF - TWF(x_t) = TF(x_t) \times TWF(x_t) \tag{5}$$

Where, TF(xt) means the word frequency of feature t in text d. TWF(xt) is a common evaluation function that is used to mark each feature and reflects the correlation between features and various types.

After the weight adjustment based on TF-TWF, the feature's importance in the classifier has changed with the weight. According to the adjusted feature's weight, modifying the feature's importance in the classifier, then we can calculate the $P(c_j \mid d)$ as follows:

$$P(c_j \mid d) = \log[P(c_j)] + \sum_{t=1}^{n} TF - TWF(x_t) \times \log[P(x_t \mid c_j)] \tag{6}$$

Where TF-TWF(xt) is a new weight function of feature xt. The feature that has a higher weight plays a greater role in the Naïve Bayesian classifier; and the feature with a smaller TF-TWF(xt) plays a smaller role in the Naïve Bayesian classifier.

Then the new decision rule of our Naïve Bayesian classifier is assigning d to the class of the maximum probability $P(c_j \mid d)$.

3.4 KNN algorithm

KNN is one of the most important non-parameter algorithms in pattern recognition field and it's a supervised learning predictable classification algorithm. The classification rules of KNN are generated by the training samples themselves without any additional data. KNN classification algorithm predicts the sample's category according to the K training samples which are the nearest neighbors to the test sample, and judge it to that category which has the largest category probability.

Suppose that there are k training categories as C1, C2, …, Cr ($1 \le r \le k$), and the number of the training samples is n. KNN algorithm categorization is as follows:

1).Convert the test sample d and training sample di into text feature vector.

2).Calculate the similarities between all training samples d and test sample di. The two commonly methods used to calculate the similarity between di and dj, which are Euclidean distance and victoria angle cosine:

$$sim(d_i, d_j) = 1 / \sqrt{\sum_{k=1}^{n} (x_{ik} - x_{jk})^2} \tag{7}$$

$$sim(d_i, d_j) = \frac{\sum_{k=1}^{n} x_{ik} x_{jk}}{\sqrt{\sum_{k=1}^{n} x_{ik}^2 \sum_{k=1}^{n} x_{jk}^2}} \tag{8}$$

3).Choose *k* samples that are bigger from *n* similarities of $sim(d_i, d_j)$. Then calculate the probability of d belong to each category respectively with the following formula.

$$P(d,C_j)=\sum sim(d_i,d_j)\bullet y(d,C_j) \quad (9)$$

Where, *y(d, C_j)* is a category attribute function, which satisfied:

$$y(d,C_j)=\begin{cases}1, d\in C_j\\ 0, d\notin C_j\end{cases} \quad (10)$$

4).Assign *d* to the category that has the largest *P(d, C_j).*

3.5 The SVM classifier

Support Vector Machines (SVMs) originate from the statistical theory and mainly solve the problem of two-class classification.

Suppose that the training set is linearly separable, |D| is the total number of members of the training set, Xi is the training samples of the feature vector i. $c_i\in\{-1,1\}$ is the category label. The linear SVM's hyper plane ($\omega\bullet x+bb=0$) and the intercept bb are determined by the following model:

$$\min\frac{1}{2}\|\omega\|^2 \quad (11)$$

$$c_i(\omega\bullet x_i+bb)\geq 1 \quad (12)$$

In fact, many samples are not classified correctly by the hyper plane. Hence, we introduce slack variables. Then, the above formula becomes:

$$\min\frac{1}{2}\|\omega\|-CC\bullet\sum_{i=1}^{|D|}\xi_i \quad (13)$$

$$c_i(\omega\bullet x_i+bb)\geq 1-\xi_i \quad (14)$$

Where, CC is a positive constant to balance the experience and confidence.

3.6 The fkNN classifier

The kNN classifier is a lazy classification algorithm. It does not need the training set. The classification effect is not ideal, especially when the classes' distribution is not uniform. Therefore, we improve the algorithm by using the fuzzy logic inference system. Its decision rule can be described as follows:

$$\mu_j(d)=\frac{\sum_{i=1}^{k}\mu_j(d_i)sim(d,d_i)\frac{1}{(1-sim(d,d_i))^{2/(b-1)}}}{\sum_{i=1}^{k}\frac{1}{(1-sim(d,d_i))^{2/(b-1)}}} \quad (15)$$

Where, *j(=1, 2, …, m)* is the number of a training set; and $\mu_j(d_i)sim(d,d_i)$ is whether sample *d* belongs to class *j* (If sample *d* belongs to class *j*, the value is *1* otherwise *0*). The fkNN decision rule is as follows:

if $\mu_j(d)=\max_i\mu_i(d)$, then $d\in c_j$.

4 Conclusion

In this paper, we present a service system, which provide services for the aging people and the only children. At the same time, the system is also an analyzing system, which collects and analyzes the vital information of the aging people by the wearable devices. All the data results about health will automatically push to the aging people or their children's mobile device. This is also a good thing for the society.

We use data mining algorithms, such as K-Means, Naïve Bayes, SVMs, and KNN, etc., to distinguish aging people from different classes. We also use the data mining algorithms to analyze the aging peoples' information. The result is a very important indicator to reflect the health status of aging people, which can be automatically pushed to aging people themselves and their children.

5 Acknowledgement

This paper is partly supported by “The comprehensive reform project of computer science and technology (ZL140103)” and partly supported by “The engineering planning project of Communication University of China (XNG1436). This paper is also supported by Guangzhou Research Institute of Communication University of China Common Construction Project, “Sunflower” – the Aging Intelligent Community.

References

[1] Ren C., Analysis on the Influence of Aging on the Healthy Development of China’s Economy and Society and Its Countermeasures, *Reform Research*, pp. 18-21, 2013.

[2] Huang R., Wang L., Research on focused crawler based on topic-related concept and page segmentation. *Application Research of Computers*, 30(8), pp. 2377-2380, 2013.

[3] Lin Z., Design and Implementation of Topic-focused Crawler. *Computer Technology and Development*, 24(8), pp. 99-102, 2014.

[4] He Y., Wang W., Xue F., Study of Massive Data Mining Based on Cloud Computing. *Computer Technology and Development*, 23(2), pp. 69-72, 2013.

[5] Huang B., Xu S., Pu W., Design and Implementation of MapReduce-Based Data Mining Platform. *Computer Engineering and Design*, 34(2), pp. 495-500, 2013.

[6] Sun X., Feng Z., Interaction Design for Wearable Devices. *Decoration*, 02(250), pp. 28-33, 2014.

[7] Zhu Y., Xu B., Wang X., Study on Technology of Intelligent Wearable Devices. *Science & Technology Information*, pp. 26-27, 2013.

[8] Shameem M.U.S., Ferdous R., An Efficient K-Means Algorithm Integrated with Jaccard Distance Measure for Document Clustering. *First Asian Himalayas International Conference*, 2009.

[9] Martin G., Christian L., Amir S.. On-line Random Naive Bayes for Tracking. *Pattern Recognition (ICPR), 2010 20th International Conference*, Istanbul , pp. 3545 – 3548, 2010.

[10] Liu H., Liu S., Su Z., An Improved KNN Text Categorization on Skew Sort Condition. *Computer Application and System Modeling (ICCASM)*, pp. 182-186, 2010.