applied sciences

MDPI

Article

Probability Analysis of Hypertension-Related Symptoms Based on XGBoost and Clustering Algorithm

Wenbing Chang [1,†], Yinglai Liu [1], Yiyong Xiao [1,†], Xingxing Xu [1], Shenghan Zhou [1,*], Xuefeng Lu [2] and Yang Cheng [3]

1 School of Reliability and System Engineering, Beihang University, Beijing 100191, China; changwenbing@263.net (W.C.); yinglailiu@buaa.edu.cn (Y.L.); xiaoyiyong@buaa.edu.cn (Y.X.); xuxx96@buaa.edu.cn (X.X.)

2 China Shipbuilding Industry Corporation No. 722 Research Institute, 430205 No. 3, Canglong Road, Jiangxia District, Wuhan 430205, China; buaasystem90@126.com

3 Center for Industrial Production, Aalborg University, Aalborg 9220, Denmark; cy@business.aau.dk

* Correspondence: zhoush@buaa.edu.cn; Tel.: +86-138-1068-2062

† These two authors contributed equally to this work.

Received: 12 January 2019; Accepted: 19 March 2019; Published: 22 March 2019

Abstract: In this paper, cluster analysis and the XGBoost method are used to analyze the related symptoms of various types of young hypertensive patients, and finally guide patients to target treatment. Hypertension is a chronic disease that is common worldwide. The incidence of it is increasing, and the age level of patients is decreasing year by year. Effective treatment of youth hypertension has become a problem in the world. In this paper, young hypertension patients are classified into two groups by cluster analysis; the proportion of different hypertension related symptoms in each group of patients is then counted; and after verifying the prediction accuracy of the XGBoost model with 10-fold cross-validation, the accuracy of clustering is calculated by the XGBoost method. The final result shows that there are significant differences in symptomatic entropy between patients with type II hypertension and those with type I hypertension. Patients with type II hypertension are more likely to have symptoms of ventricular hypertrophy and microalbuminuria. Through this analysis, patients can have preventive treatment according to their own situation, and this can reduce the burden of medical expenses and prevent major diseases. Applying the data analysis into the medical field has great practical significance.

Keywords: hypertension; cluster analysis; XGBoost algorithm; hypertension related symptoms

1. Introduction

The world economy is developing rapidly in recent decades, and cardiovascular disease has become the main cause of death in the world [1]. According to the survey, at present, 30% of the world's deaths are caused by 52 cardiovascular diseases [2]; hypertension is a disease associated with cardiovascular disease, accounting for 13% of cardiovascular diseases, that is, the high prevalence of hypertension worldwide is one of the reasons for the high prevalence of cardiovascular diseases. It is a chronic disease and has a serious impact on human health. Over time, its prevalence worldwide has increased year by year, but its awareness rate and control rate are still very low. Hypertension can damage the vascular system, causing major diseases such as myocardial infarction, coronary heart disease and stroke secondary diseases. Studies by L Gray and IM Lee [3] have shown that if blood pressure is high in early adulthood, the mortality rate of cardiovascular disease and coronary heart disease will increase in a few decades. Around 54% of the world's strokes, 47% of ischemic heart

disease, 75% of hypertension, and 25% of other cardiovascular diseases are caused by high blood pressure [4].

In China, the number of people suffering from hypertension is on the rise due to poor awareness, treatment and control of hypertension, which is much lower than that in the West, especially among young people and people in remote areas. The incidence of hypertension has increased greatly. In 1991, the number of people over 15 years of age suffering from hypertension increased by 60% compared with 1980. In 2002, the prevalence rate of hypertension in those over 18 years old reached 18.8%. Two out of 10 people suffered from hypertension, and the number with high blood pressure reached 153 million, accounting for 20% of the world's total hypertension. Mild to moderate hypertension accounts for 90% of cases and normal blood pressure does not exceed 50% [5]. According to a recent survey, about one-third of adults suffer from hypertension. Half of them will receive treatment for hypertension, but only 5% of them will eventually have their blood pressure properly controlled [6]. In China 750,000 people die of hypertension every year. Obesity is a major contributor to hypertension. In China, the proportion of obese and overweight people aged 7–18 increased 28-fold between 1985 and 2000. In 1992–2002, 260 million people were overweight or obese, so hypertension will continue to deteriorate in the future [7]. Research on hypertension has become a key issue. Hypertension itself is not terrible, but it can cause enormous damage to the target organs. The target organs of hypertension are mainly the heart, kidney, brain, blood vessels and so on. Tatasciore et al. [8], O'Sullivan et al. [9], and Gao et al. [10] studied the damage of target organs caused by hypertension, and it turned out that high blood pressure causes great damage to the target organs.

Hypertension has become a very common disease worldwide, so in view of its great harm to the human body, how to detect hypertension and treat it is urgent. The classification of hypertension in Chinese hospitals is only based on severity grades, such as hypertension grade 1 and hypertension grade 2. According to the severity of the patients, appropriate drugs are used to treat hypertension. Although this classification method can be applied, it is not rigorous. As mentioned above, half of hypertensive patients will suffer from target organ damage; in some it caused kidney damage, in some it caused heart damage, some people suffer damage to one target organ and others to two target organs, which means that although these patients are judged to have hypertension, there are differences between them. Patients with the same target organ injury may have some similarities, which requires analysis of their detection indicators data, which cannot be generalized, so the current hospital classification according to the severity of the disease is not rigorous. For most Chinese families, the cost of medical care is a large burden. If a patient can be detected with hypertension, at the same time according to various indicators it can be determined which type of hypertension he has and which target organs may be damaged due to hypertension in the future. This can avoid the blindness of treatment, and doctors can target preventive treatment to patients. As time goes by, more patients with hypertension will be free from the trouble of hypertension target organ damage.

Most of the current studies are based on the analysis of the effects of a certain physical index or a certain feature of target organ damage in hypertensive patients, according to Viazzi F et al. [11], Giuseppe Mule et al. [12], but there is almost no prediction of hypertension-related symptoms. With the continuous development of technology, the era of big data has arrived. Applying big data analysis methods to medical, education, and smart cities has become a global concern [13,14]. The hospital generates a large amount of body index information at all times, so it is necessary to apply the big data analysis method to the medical field. Through effective analysis and processing, the efficiency and accuracy of diagnosis can be improved, and then scientific and accurate treatment methods can be used to treat patients.

The innovation of this paper is that, according to the patient's detection index, the blood pressure category of a patient can be obtained by the clustering algorithm, and then combined with the XGBoost method, the probability that the patient has different hypertension-related symptoms can be obtained. For a new patient, after identifying the category by the clustering algorithm, the doctor can predict the

possible symptoms based on the results obtained in this paper. This paper combines the hospital's big data and algorithms to solve medical problems, which is of great practical significance.

The content of this article is arranged as follows: the first chapter is about the current situation of hypertension in China, the introduction of hypertension research and the innovation of this article. The second chapter introduces the main algorithms used in this paper, including clustering algorithm, XGBoost method and the overall method architecture used in this paper. The third chapter is the analysis of the experimental process and the analysis of the results. The fourth chapter summarizes this article.

2. Materials and Methods

2.1. Clustering Methodology

Cluster analysis is based on the similarity of each class. The similarity between one point and other points in a class is greater than the similarity with other classes. The clustering method can maximize the similarity of objects in the class, and the similarity between classes is the smallest [15]. K-means is a kind of clustering technology applied to Web, which randomly selects clustering centers in all data and classifies them according to the distance between data points and centers [16]. The main idea is to find the k class centers of the data set and classify the data into k classes so that the distance between the data points in the data set and the class centers of the classes to which they belong is minimal. This method often requires manual determination of the number of clustering centers, which is subjective. When the number of classes selected manually does not match the data, there will be significant errors [17–19]. Hierarchical clustering [20], density-based [21] are also commonly used clustering methods.

The k-means method mentioned above is the most basic method among clustering methods [22], because it relies too much on the initial selection of k aggregation centers, resulting in a series of grid-based clustering methods and so on. This article uses a new clustering method proposed by Alex Rodriguez and Alessandro Laio [23].

The method has two distinct characteristics as follows:

The density of the cluster center is large, and its density around it is smaller.

The distance between the cluster center and other dense data points is relatively far.

$S = \{x_i\}_{i=1}^{N}$ represents data sets with clustering. $I_s = \{1, 2, \cdots, N\}$ is the corresponding index set. The distance between x_i and x_j represented by $d_{ij} = dist(\boldsymbol{X}_i, \boldsymbol{X}_j)$. For every point x_i in S, ρ_{i} and δ_{i} representing the two characteristics mentioned above where ρ_{i} represents local density, δ_{i} represents the distance (Formula (3) is explained in detail). Here are two calculation methods for calculating local density.

Cut-off kernel:

$$\rho_i = \sum_{j \in I_s \setminus \{i\}} \chi(d_{ij} - d_c) \tag{1}$$

Where function:

$$\chi(x) = \begin{cases} 1, & x < 0; \\ 0, & x \geq 0; \end{cases}$$

Among them, d_c represents the truncation distance. This is an amount given by experience and requires artificial settings. When choosing d_c, it should be emphasized that there should be a number of 1%–2% of the total number of data points around each data point. For each data point in $S = \{x_i\}_{i=1}^{N}$, it has a distance value from the point of $N-1$ other than itself, so there are $N(N-1)$ distances in $S = \{x_i\}_{i=1}^{N}$. But the distance between any two points is computed twice, so the number of effective distances is $M = \frac{1}{2}N(N-1)$. We sort the distance D_{ij} from large to small, and the sequence is $d_1 \leq d_2 \leq \cdots \leq d_M$, Take d_c as $d_k, k = \{1, 2, \cdots M\}$. In general, the distance satisfying the condition of less than d_c takes up about $\frac{k}{m}$, that is, about $\frac{k}{m}N(N-1)$, and the average distance to each data point is about $\frac{k}{m}N(N-1) = \frac{k}{m}N$. The ratio $\frac{k}{m}$ is equivalent to t in the algorithm. So when the ratio t is

given, k can be approximately taken as Mt. In this paper, the range of t is locked at 1~2%, which is based on the empirical values of many data sets.

ρ_i represents the data points between S and x_i (except x_i itself) with a distance less than d_c. Gaussian kernel:

$$\rho_i = \sum_{j \in I_s \backslash \{i\}} e^{-\left(\frac{d_{ij}}{d_c}\right)^2} \tag{2}$$

Compared with the Formula (1), the cut-off kernel is a discrete value, but the Gaussian kernel is a continuous value, so the probability that different data points have the same local density is smaller under the latter condition, and the latter also satisfies the property that the closer to the data point, the greater the density.

Distance δ_{i}:

Assuming that $\{q_i\}_{i=1}^N$ represents a descending order of $\{\rho_i\}_{i=1}^N$, it means that it satisfies:

$$\rho_{q1} \geq \rho_{q2} \geq \cdots \geq \rho_{qN}$$

Defined as follows:

$$\delta_{q_i} = \begin{cases} \min\limits_{\substack{q_i \\ j<i}} \{d_{q_i q_j}\}, & i \geq 2; \\ \max\limits_{j \geq 2} \{\delta_{q_i}\}, & i = 1. \end{cases} \tag{3}$$

According to the Formula (3), when the local density of x_i is the largest. δ_{i} represents the distance of the point in S where the distance from x_i is the largest, otherwise δ_{i} represents the distance from the point where the distance from x_i is the smallest among the points where the local density is larger than x_i. So far, for each data point x_i in S, we can calculate $(\rho_{\mathrm{i}}, \delta_i), i \in I_s$. For example, in a two-dimensional space, 20 points are distributed, and the numbers in the circles represent the number of points, and the distribution of points is shown in Figure 1A. The ρ and δ of each point are calculated and represented in a two-dimensional space, as shown in Figure 1B. ρ is the horizontal axis and δ is the longitudinal axis. Figure 1B plays a decisive role in the selection of clustering centers, because subjective selection plays a great role in the selection of clustering centers, it is necessary to create a new metric γ, the product of ρ and δ, the greater the product, the higher the probability of being selected as a clustering center.

$$\gamma_i = \rho_i \delta_i, i \epsilon Is \tag{4}$$

The algorithm flow chart of this method is as follows.

Step 1: Initialization and Preprocessing

- Given a parameter $t \in (0, 1)$ which is the same as determining the cut-off distance d_c
- Calculate the distance d_{ij}, and make $d_{ij} = d_{ji}, i < j, i, j \in I_s$
- Determining cut-off distance d_c
- Compute $\{\rho_i\}_{i=1}^N$ and generate its descending order subscript $\{q_i\}_{i=1}^N$
- Calculate $\{\delta_i\}_{i=1}^N$ and $\{n_i\}_{i=1}^N$

Step 2: Select the cluster center $\{m_j\}_{j=1}^{n_c}$ and initialize the data point classification attribute tag $\{c_j\}_{i=1}^n$, as follows:

$$c_i = \begin{cases} k, & \text{If } \mathrm{x_i} \text{ is the distance center, it belongs to the first k cluster.} \\ -1, & other \end{cases}$$

Step 3: Categorization of non-clustered central data points

Step 4: If $n_c > 1$, the data points in each cluster are further divided into cluster core and cluster halo

- Initialization mark $h_i = 0, i \in I_S$
- Generate an average local density upper bound $\left\{\rho_i^b\right\}_{i=1}^{n_c}$ for each cluster.
- Identification cluster halo

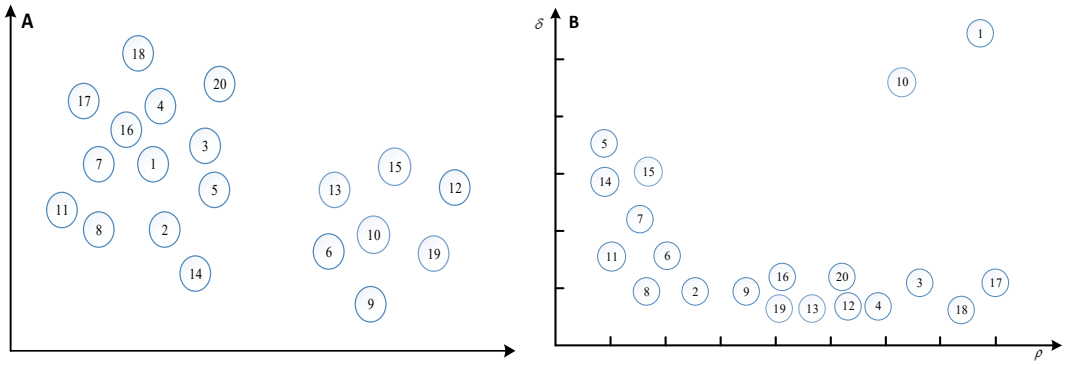


Figure 1. Example (**A**) Distribution of random points; (**B**) The ρ and δ values of each point.

2.2. XGBoost Method

The full name of XGBoost is eXtreme Gradient Boosting, which is a c++ implementation of Gradient Boosting Machine algorithm. The greatest advantage of this algorithm is that it can automatically use the CPU to run multi-threaded, improve efficiency, and the algorithm has been changed to improve the accuracy. Shi et al. [24], Lei et al. [25], Zhang et al. [26] used the XGboost method to obtain very accurate predictions. This article uses the XGBoost method proposed by Chen T Q and Guestrin C [27].

The algorithm steps can be expressed as follows:

Objective function:

$$\begin{gathered} \zeta(\phi) = \sum_{\mathrm{i}} l(\hat{y}_i, y_i) + \sum_{\mathrm{k}} \Omega(f_k) \\ \text{where } \Omega(f) = \gamma T + \tfrac{1}{2}\lambda \|w\|^2 \end{gathered} \tag{5}$$

Training objective function:

$$\zeta^{(\mathrm{t})} = \sum_{i=1}^{n} l\left(y_i, \hat{y}^{(t-1)} + f_t(x_i)\right) + \Omega(f_t) \tag{6}$$

Target function Taylor second-order expansion approximation:

$$\begin{gathered} \zeta^{(t)} \cong \sum_{i=1}^{n} \left[l\left(y_i, \hat{y}^{(t-1)}\right) + g_i f_t(x_i) + \tfrac{1}{2} h_i {f_t}^2(x_i) \right] + \Omega(f_t) \\ \text{where } \mathrm{g_i} = \delta_{\hat{y}^{(t-1)}} l\left(y_i, \hat{y}^{(t-1)}\right) \text{ and } h_i = \delta^2_{\hat{y}^{(t-1)}} l\left(y_i, \hat{y}^{(t-1)}\right) \end{gathered} \tag{7}$$

Remove the constant term:

$$\zeta^{(t)} = \sum_{i=1}^{n} \left[g_i f_t(x_i) + \frac{1}{2} h_i {f_t}^2(x_i) \right] + \Omega(f_t) \tag{8}$$

Find the optimal solution of the objective function:

$$\overline{\zeta}^{(\mathrm{t})}(q) = -\frac{1}{2} \sum_{j=1}^{T} \frac{\left(\sum\limits_{i \in g_i} g_i \right)^2}{\sum\limits_{i \in h_i} h_i + \lambda} + \gamma T \tag{9}$$

XGBoost has the following advantages:

1. It supports linear classifier.
2. In order to control the complexity of the model, the XGboost method deliberately sets a regularization term in the objective function.
3. XGBoost uses the second-order Taylor expansion of the objective function, and uses the first and second derivatives.
4. When tree nodes split, we need to calculate the corresponding gain of each segmentation point of each feature, that is, to enumerate all possible segmentation points with a greedy method.
5. XGBoost borrows from the random forest approach and supports column sampling, which not only reduces over-fitting, but also reduces computational complexity.
6. The XGBoost method considers the problem of efficient use of disks, especially when the amount of data is large and there is not enough memory. It combines data compression and fragmentation, and a multi-threaded approach that greatly improves efficiency.
7. Shrinkage is equivalent to learning rate. After an iteration, XGBoost multiplies the weight of the leaf nodes by this coefficient, mainly to weaken the impact of each tree, so that there is more learning space. In practical applications, eta is generally set up a little bit, and then the number of iterations is set a little larger.

2.3. Probability Analysis Process

The flow chart of the probability analysis is shown in Figure 2. First of all, we get the original data of hypertensive patients for physical examination, because not all indicators are related to hypertensive diseases, so we need to pick out the data related to hypertension from the original data. Not all patients have done all the necessary indicators related to hypertension, so there will be many missing values in the original data. In order to make the calculation more accurate, it is necessary to delete the missing samples. There are many indicators related to hypertension, and the differences among them are quite large. Some of them are very small, in the range of 0–10, and some are very large, and may reach more than 100. Therefore, data need to be standardized, and the target value will be locked within 0–1.

There are many indicators related to hypertension. If all these indicators are used, it will not only increase the huge computational burden, but also make the calculation very difficult. In addition, in most cases, there is a certain correlation between variables, which leads to the overlap of these variables in the information of the problem. In order to minimize the number of variables in this paper, principal component analysis (PCA) is used. The main idea of this method is to maximize the possibility of deleting some extra or highly relevant variables, and then create as few new irrelevant variables as possible. By PCA, several principal components will be selected according to the cumulative contribution rate, and then the scores $x_1, x_2, \cdots x_n$ of several principal components will be obtained. The scores $(x_1, x_2, \cdots x_n)$ of principal components will be taken as the coordinates of this point. According to the coordinates, the distance between every two points will be calculated. The feature information of the original sample is processed by the PCA to obtain the coordinates of the point, so each point uniquely represents a sample.

Through the distance between points, the patients are clustered by clustering algorithm to get the categories of each patient, and then each patient is labeled according to the clustering results. According to the categories of patients and their real symptoms, the hypertension-related symptoms of different types of hypertensive patients were counted separately, and the true proportion of each symptom was obtained.

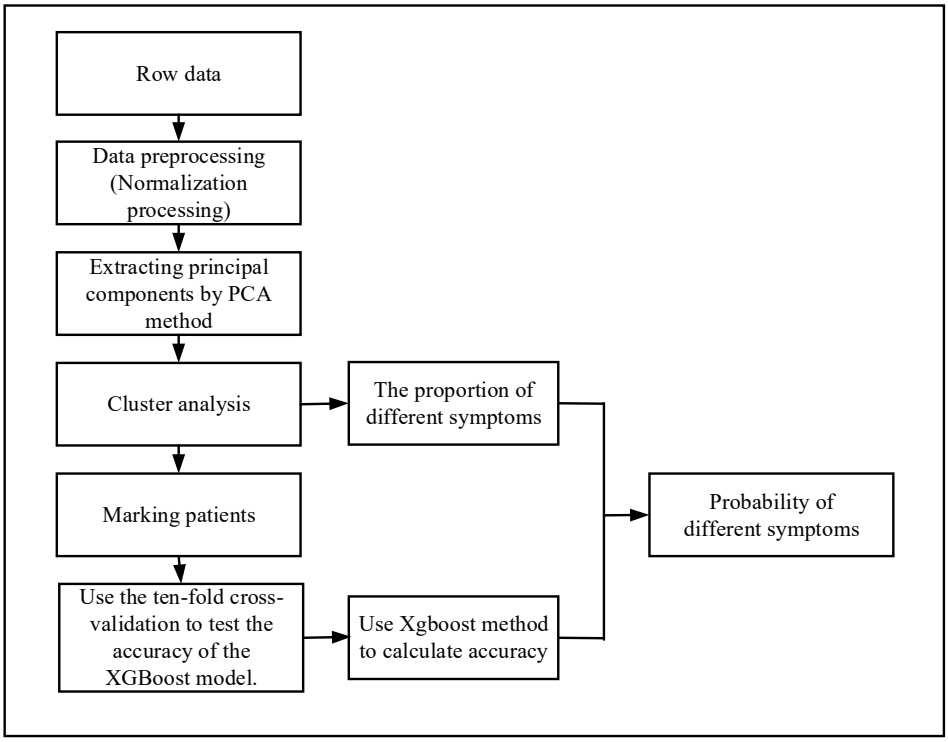


Figure 2. Probability analysis flow chart.

In order to calculate the accuracy of the clustering effect, the XGBoost algorithm is used. Firstly, a part of the samples is selected as the training set, and then the rest is used as the test set to calculate the accuracy. A very important point here requires special attention, in order to determine that the XGBoost method can accurately calculate the accuracy of cluster analysis, the prediction performance of the XGBoost model should be verified. Therefore, the performance of XGBoost is estimated using a 10-fold cross-validation method. If XGBoost works well, it can be used to calculate the accuracy of the clustering method. After clustering, the proportion of patients with different symptoms in each group has been calculated, multiplying the proportion by the accuracy of the XGBoost method yields the probability that a patient will have a certain symptom. The above are the steps of the diagnostic model.

3. Experiment and Result Discussion

The data used in this study was obtained from 531 hypertensive patients in a hospital in Beijing. A total of 22 related indicators of blood pressure were used. Hypertension-related indicators are shown in Table 1.

Table 1. Hypertension related indicators.

Number	Indicator	Number	Indicator
1	RARMSBP	12	ABIL
2	RARMDBP	13	MEANSBP
3	LARMSBP	14	MEANDBP
4	MARMDBP	15	HIGHSBP
5	RLEGSBP	16	HIGHDBP
6	RLEGDBP	17	LOWSBP
7	LLEGSBP	18	LOWDBP
8	LLEGDBP	19	DAYMSBP
9	BAPWVR	20	DAYMDBP
10	BAPWVL	21	NIHTMSBP
11	ABIR	22	NIHTMDBP

RARMSBP is right upper limb systolic pressure; RARMDBP is right upper limb diastolic pressure; LARMSBP is left upper limb systolic pressure; LARMDBP is left upper limb diastolic pressure; RLEGSBP is right lower limb systolic pressure; RLEGDBP is right lower limb diastolic pressure; LLEGSBP is left lower limb diastolic pressure; BAPWVR, ABIR, ABIL are all one of limb blood pressure. MEANSBP is the 24-h mean systolic blood pressure; MEANDBP is the 24-h mean diastolic blood pressure; HIGHSBP is the highest systolic blood pressure; HIGHDBP is the highest diastolic blood pressure; LOWSBP is the lowest systolic blood pressure; LOWDBP is the lowest diastolic blood pressure. DAYMSBP is the average systolic blood pressure during the day; DAYMDBP is the average diastolic blood pressure during the day; NIHTMSBP is the average systolic blood pressure at night; and NIHTMDBP is the average diastolic blood pressure at night.

Because of the relevant indicators obtained by physical examination, the parameter gap of different people is relatively large, which is not conducive to the later research. Therefore, the index parameters should be normalized firstly, and the values of each index should be locked in the range of 0 to 1.

In this analysis, there are 22 variables related to hypertension. Too many variables will increase the complexity of the problem. In reality, many variables are related, which means the relevant variables have redundant information. In view of this, the principal component analysis method is adopted. Under the premise of retaining the original information to the greatest extent, the variables with large correlations in the original variables are deleted, and several unrelated new variables are constructed.

The contribution of 22 variables to the variance is shown in Table 2. The table shows the contribution rate of each variable to the original information and the cumulative contribution rate of all variables. Two new variables are selected to describe the original problem by principal component analysis (PCA), because the two new variables can keep 75% of the original information.

Table 2. Contribution table.

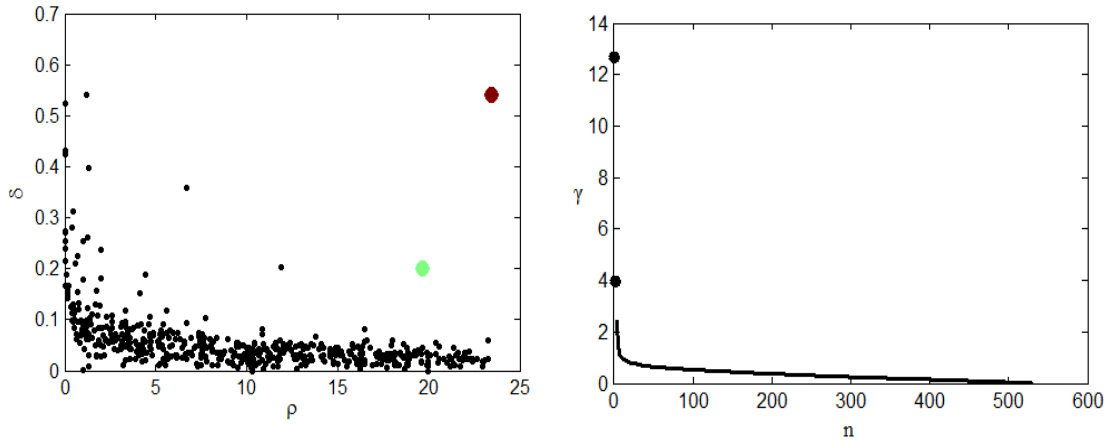
Variable Number	Component Contribution	Cumulative Contribution	Variable Number	Component Contribution	Cumulative Contribution
1	0.57	0.57	12	0.01	0.98
2	0.18	0.75	13	0.00	0.99
3	0.05	0.80	14	0.00	0.99
4	0.05	0.85	15	0.00	0.99
5	0.04	0.89	16	0.00	0.99
6	0.03	0.92	17	0.00	1.00
7	0.02	0.94	18	0.00	1.00
8	0.01	0.95	19	0.00	1.00
9	0.01	0.96	20	0.00	1.00
10	0.01	0.97	21	0.00	1.00
11	0.01	0.98	22	0.00	1.00

The scores of two new variables Y_1, Y_2, Y_1 and Y_2 were obtained on the basis of the original 22 variables. The number of elements contained in Y_1 and Y_2 was 531, that is, the number of patients. (Y_{1i}, Y_{2i}) as the coordinates of point i, Results are shown in Table 3. The distance between each point can be calculated according to the Euclidean distance formula. The result of the final calculation is a 531×531 matrix. Here the distance of oneself to oneself is set to 0.

Table 3. Principal component score (coordinates).

Number	Y_{1i}	Y_{2i}	Number	Y_{1i}	Y_{2i}
1	0.232718	−0.04398	8	0.121475	0.212878
2	−0.08059	0.067648	9	−0.01286	−0.39306
3	0.137897	0.203493	10	0.94541	−0.07324
4	0.987187	−0.21345	11	1.147777	−0.14306
5	−0.22769	−0.05435	...	...	...
6	−0.81281	0.060526	530	0.312253	0.048231
7	0.463736	0.083148	531	−0.3606	0.026286

Figure 3A shows the clustering results. The red and green points in the graph represent the clustering centers of two types. This graph shows the relationship between the density and distance of each point. The clustering centers usually fall on the right side of the graph. Two black-colored points in Figure 3B. represent two clustering centers, and it is obvious that the γ values of these two points are large, so they will be given priority when considering clustering centers. After the center point is determined, points other than the two center points must select a class based on the conditions of the cluster.



(**A**) (**B**)

Figure 3. Clustering diagram. (**A**) The ρ and δ values of each point; (**B**) The γ values of each point.

The final clustering results are shown in Table 4. The total sample size is 531. The number of patients in the first category is 138, and the center point is the patient numbered 217. In a class, the average density is calculated, the point where the density exceeds the average density is at the core position, and the point where the density is lower than the average density is due to the position of the halo. According to the density of the sample points, 13 people are in the core position of the first category, and 125 people are in the halo position. The number of patients in the second category is 393, and the center point is the number 355. Similarly, according to the density of the sample points, 64 people are in the core position of the second category, and 329 people are in the halo position.

Table 4. Clustering result.

Category	Core	Halo	Total	Center
1	13	125	138	217
2	64	329	393	355

A represents ventricular hypertrophy, B represents vascular sclerosis, C represents lower limb ischemia, D represents renal insufficiency, E represents microalbuminuria, F represents fundus disease, G represents stroke.

Table 5 describes an analysis of the related symptoms of different types of hypertension after classification. There are two types of hypertension, the number of one group of patients was 138, and the number of two groups of patients was 393. The table shows the number of patients with different types of hypertension corresponding to different symptoms, and the corresponding proportion of each type of symptoms, the appearance of different symptoms means that there is damage to the corresponding target organ, such as ventricular hypertrophy, and shows that hypertension caused damage to the heart of patients and that renal insufficiency means that hypertension caused damage to the kidneys of patients. For heart target organs, only about 10% of the patients in the first group developed ventricular hypertrophy, while 34.9% of the patients in the second group developed ventricular hypertrophy. For vascular sclerosis, both the first group and the second group had a very high proportion of the symptoms; the first group was close to 90% of the patients, the second group of patients is as high as 96.4% of the proportion, which means patients with hypertension will generally appear in the symptoms of vascular sclerosis. For lower limb ischemia, the proportion of patients in both groups is not high, 2.2% and 0.8% respectively, which proves that patients with hypertension rarely have symptoms of lower limb ischemia. Similarly, for renal insufficiency, the proportion of minor symptoms in both groups was small, but the proportion of patients in the second group was six times higher than that in the first group. For microalbuminuria, 41.2% of patients in the second group had this symptom, compared with 7.2% in the first group. Fundus lesions and stroke were not significantly different between the two groups, and the incidence is very low. The results suggest that there is indeed a significant difference in some symptoms between different types of hypertension.

Table 5. Symptom analysis.

Category	1(138)		2(393)	
Symptom	**Number of People**	**Ratio**	**Number of People**	**Ratio**
A	14	10.1%	137	34.9%
B	123	89.1%	379	96.4%
C	3	2.2%	3	0.8%
D	1	0.7%	16	4.1%
E	10	7.2%	162	41.2%
F	1	0.7%	13	3.3%
G	8	5.8%	25	6.4%

After cluster analysis of 531 patients, the patients were divided into two groups and the symptoms with higher prevalence were selected. The proportion of patients with vascular sclerosis in the first group was 89.1%, that of patients with vascular sclerosis in the second group was 96.4%, that of patients with ventricular hypertrophy was 34.9%, and that of microalbuminuria was 41.2%.

In order to determine the accuracy of clustering, 334 of 531 patients were trained and 197 patients were tested by XGBoost method, the ratio of the test set is 37%. In the XGBoost algorithm, the depth of the tree is 6 and the total number of iterations, i.e., the number of decision trees is 100; the gbtree tree model is used as the base classifier; all CPU parallel computing is used; and each tree is trained, using the full training set, with the other parameters the default values. In order to estimate the XGBoost prediction performance, the 10-fold cross-validation method is used here. The accuracy of each cross-validation is shown in Table 6, n is the nth verification of the ten-fold cross-validation. The average of the 10 accuracy rates is used as the index of XGBoost's prediction performance. The final result is 0.973, indicating that the XGBoost method has good prediction performance and it can be used to calculate the accuracy of the clustering effect.

Table 6. Ten-fold cross-validation effect evaluation.

n	Accurate	*n*	Accurate
1	1.000	6	0.970
2	1.000	7	0.970
3	0.941	8	1.000
4	0.971	9	0.939
5	0.941	10	1.000

After determining that the XGBoost method can be used to calculate the accuracy of the clustering effect, the result of the XGBoost method is 98.48%. This means that there is a probability of 98.48% that the result of the clustering algorithm is correct, showing that the result of clustering analysis is very high, and there is no obvious error in general.

Figure 4 shows the importance degree of the features obtained by the XGBoost method. It can be seen from the figure that the 13th and 20th features have a great influence on the results. According to Table 1, the two variables are the 24-h average systolic blood pressure and systolic blood pressure during the day.

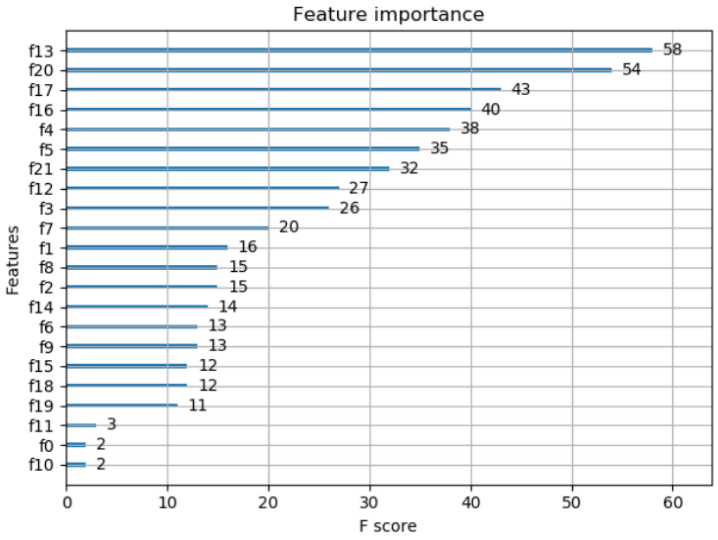


Figure 4. Feature importance.

XGBoost is a boosting tree method in which each decision tree can be drawn. Figure 5 is the decision tree obtained by this model. It can be seen that the root node is the 13th feature, and the left subtree of the second layer is the 20th feature, which is consistent with the feature importance obtained in Figure 4.

After clustering patients, the proportion of various symptoms in different categories can be counted, this ratio is multiplied by the accuracy obtained by the XGBoost method, and the probability of occurrence of various symptoms is obtained, which is shown in Table 7. This method of calculation is similar to the method of calculating the expectation in terms of probability.

According to Table 7, if a new patient is identified as the first type of hypertension, then the probability that he has arteriosclerosis is 87.75%. If he is identified as the second type of hypertension, his probability of arteriosclerosis is 94.93%. These two values are very high, indicating that if the patient is diagnosed with hypertension, then the patient is likely to suffer from vascular sclerosis, patients can receive hardening therapy immediately. According to the data from the table, the difference between ventricular hypertrophy and microalbuminuria is obvious. That is to say, if hypertensive patients are classified into the second group, they will have a greater chance of developing symptoms of ventricular hypertrophy and microalbuminuria, and they can be treated preventively.

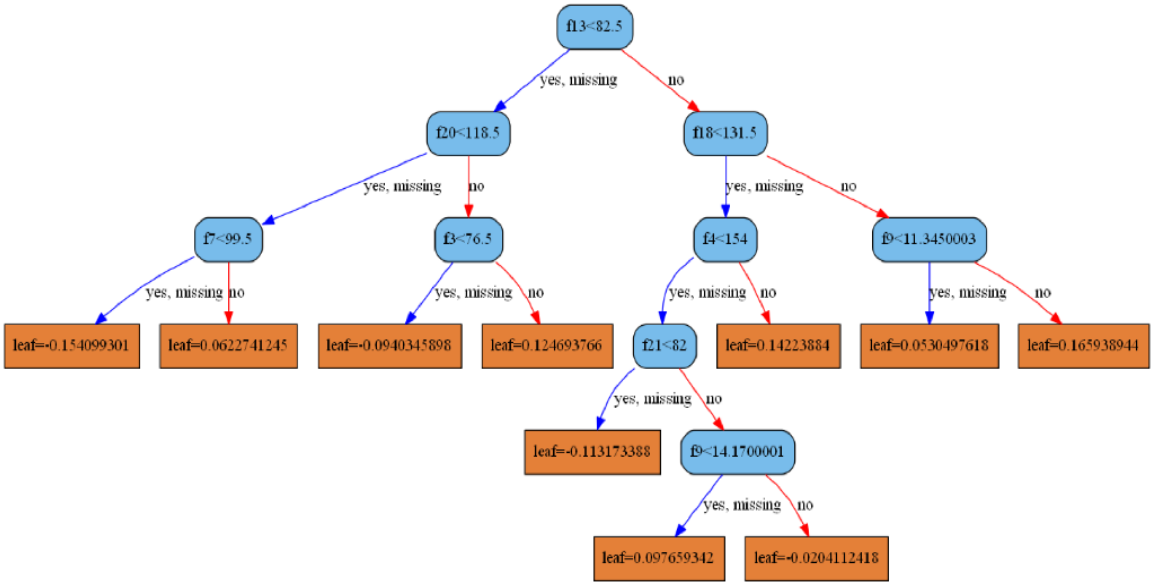


Figure 5. XGboost node graph.

Table 7. Clustering and XGBoost methods for predicting symptom probability summary table.

Category	**1**		**2**	
XGBoost Prediction Result		**98.48%**		
Symptom	**Ratio**	**Probability**	**Ratio**	**Probability**
Ventricular Hypertrophy	10.10%	9.95%	34.90%	34.37%
Vascular Sclerosis	89.10%	87.75%	96.40%	94.93%
Lower Limb Ischemia	2.20%	2.17%	0.80%	0.79%
Renal Insufficiency	0.70%	0.69%	4.10%	4.04%
Microalbuminuria	7.20%	7.09%	41.20%	40.57%
Fundus Disease	0.70%	0.69%	3.30%	3.25%
Stroke	5.80%	5.71%	6.40%	6.30%

4. Conclusions

In this paper, we first pretreated hypertensive data and extracted principal components. After labeling patients with clustering algorithm, combined with XGBoost, the probability of different symptoms of patients was calculated. The final results show that hypertension patients have a high probability of vascular sclerosis symptoms. In other symptoms, there are significant differences between the two groups. The second group has a greater probability of ventricular hypertrophy and microalbuminuria symptoms than the first group. Therefore, if a hypertensive patient is divided into the second type of hypertensive patients, we can focus on the prevention and treatment of microalbuminuria and ventricular hypertrophy. Hypertension can damage organs, which is painful for patients. Microalbumin is the most sensitive and reliable diagnostic indicator for early detection of nephropathy, and ventricular hypertrophy is a precursor of heart disease. Therefore, preventive treatment of type II hypertension patients can reduce the damage of hypertension to the kidney and heart.

For a new patient, classification can be determined by the clustering method in this paper. If many patient categories have been determined by this clustering method, these patients can be added as sample points to the original sample and the probability of different symptoms recalculated. As time goes by, the sample will continue to increase and the final probability will be more accurate.

At present, young people with hypertension in China are increasing year by year, and obesity is becoming more common among young people, meaning that within a short time the condition of hypertension in China will further deteriorate. Hypertension itself is not a major threat, but it will

cause damage to target organs. For patients, if the target organ damage occurs, it will not only have a great financial burden, but also have a great impact on the body and spirit. In conclusion, the analysis of hypertension-related symptoms and target organ damage is urgent.

The method proposed in this paper can obtain the probability of certain symptoms in different types of hypertensive patients, and it can guide different types of hypertensive patients for targeted treatment, greatly reducing the risk of suffering from more serious diseases, with strong practical significance.

With the continuous advancement of the era of big data and the continuous development of the medical field, more and more medical data will be preserved. Utilizing big data analysis methods to scientifically use these saved data to increase the accuracy of medical diagnosis will be an important research direction in the future.

This paper combines big data analysis methods with a hospital's data, and successfully applies big data analysis methods to the medical field. This method can predict the probability of some related symptoms in hypertensive patients. In the medical field, big data analysis can play a huge role, but the extensive application of data analysis in the medical field requires a lot of research. With the continuous development of big data analysis, more accurate research results will make outstanding contributions in the medical field.

Author Contributions: This manuscript was written by Y.L under the supervision of W.C., Y.X and S.Z. The modeling, data analysis and software process was executed by Y.L., X.X, X.L and Y.C are responsible for data acquisition and model design.

Funding: This work is supported by the National Natural Science Foundation of China (Grant No. 71501007 and 71672006 and 71871003). The study is also sponsored by the Aviation Science Foundation of China (2017ZG51081), the Technical Research Foundation (JSZL2016601A004).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Hou, D.Z.; Chen, J.; Ren, X. A whole foxtail millet diet reduces blood pressure in subjects with mild hypertension. *J. Cereal Sci.* **2018**, *84*, 13–19. [CrossRef]
2. Kearney, P.M.; Whelton, M. Global burden of hypertension: Analysis of worldwide data. *Lancet* **2005**, *365*, 217–223. [CrossRef]
3. Gray, L.; Lee, I.M.; Sesso, H.D. Blood pressure in early adulthood, hypertension in middle age, and future cardiovascular disease mortality: HAHS (Harvard Alumni Health Study). *J. Am. Coll. Cardiol.* **2011**, *58*, 2396–2403. [CrossRef] [PubMed]
4. Lawes, C.M.; Vander, H.S. Global burden of blood-pressure-related disease, 2001. *Lancet* **2008**, *371*, 1513–1518. [CrossRef]
5. Liu, L.S. Chinese guidelines for the management of hypertension. *Chin. J. Hypertens.* **2011**, *39*, 579–615.
6. Lewington, S.; Lacey, B.; Clarke, R. The Burden of Hypertension and Associated Risk for Cardiovascular Mortality in China. *JAMA Intern. Med.* **2016**, *176*, 524. [CrossRef] [PubMed]
7. Gao, Y.; Chen, G.; Tian, H. Prevalence of hypertension in China: A cross-sectional study. *PLoS ONE* **2013**, *8*, e65938. [CrossRef] [PubMed]
8. Tatasciore, A.; Renda, G.; Zimarino, M. Awake systolic blood pressure variability correlates with target-organ damage in hypertensive subjects. *Hypertension* **2007**, *50*, 325–332. [CrossRef] [PubMed]
9. O'Sullivan, C.; Duggan, J.; Lyons, S. Hypertensive target-organ damage in the very elderly. *Hypertension* **2003**, *42*, 130–135. [CrossRef]
10. Gao, H.; Yan, Y. Analysis of Target Organ Damage in Youth with Pre-hypertension. *Chin. J. Rehabil. Theory Pract.* **2012**, *18*, 760–762.
11. Viazzi, F.; Parodi, D.; Leoncini, G. Serum uric acid and target organ damage in primary hypertension. *Hypertension* **2005**, *45*, 991–996. [CrossRef]
12. Giuseppe, M.; Nardi, E.; Cottone, S. Influence of metabolic syndrome on hypertension-related target organ damage. *J. Intern. Med.* **2005**, *18*, 503–513.

13. Lytras, M.D.; Raghavan, V.; Damiani, E. Big data and data analytics research: From metaphors to value space for collective wisdom in human decision making and smart machines. *Int. J. Semant. Web Inf. Syst.* **2017**, *13*, 1–10. [CrossRef]
14. Lytras, M.D.; Aljohani, N.R.; Hussain, A.; Luo, J.; Zhang, X.Z. Cognitive Computing Track Chairs' Welcome & Organization. In Proceedings of the Companion of the Web Conference, Lyon, France, 23–27 April 2018.
15. Jin, J.G. Review of Clustering Method. *Comput. Sci.* **2014**, *41*, 288–293.
16. Burney, S.M.A.; Tariq, H. K-Means Cluster Analysis for Image Segmentation. *Int. J. Comput. Appl.* **2014**, *96*, 1–8.
17. Celebi, M.E.; Kingravi, H.A.; Vela, P.A. comparative study of efficient initialization methods for the k-means clustering algorithm. *Exp. Syst. Appl.* **2013**, *40*, 200–210. [CrossRef]
18. Coates, A.; Ng, A.Y. Learning Feature Representations with K-Means. *Lect. Notes Comput. Sci.* **2012**, *7700*, 561–580.
19. Zhong, L.; Tang, K.; Lin, L. An improved clustering algorithm of tunnel monitoring data for cloud computing. *Sci. World J.* **2014**, *2014*, 630986. [CrossRef] [PubMed]
20. Barjoseph, Z.; Gifford, D.K.; Jaakkola, T.S. Fast optimal leaf ordering for hierarchical clustering. *Bioinformatics* **2001**, *17*, S22–S29. [CrossRef]
21. Zhi, Q.B.; An, X. Border-Processing Technique in Grid-Based Clustering. *Patt. Recognit. Artif. Intell.* **2006**, *19*, 277–280.
22. Chang, W.B.; Xu, Z.Z.; Zhou, S.H. Research on detection methods based on Doc2vec abnormal comments. *Fut. Gener. Comput. Syst.* **2018**, *86*, 656–662. [CrossRef]
23. Rodriguez, A.; Laio, A. Machine learning. Clustering by fast search and find of density peaks. *Science* **2014**, *344*, 1492. [CrossRef] [PubMed]
24. Shi, T. XGBoost-based enterprise failure risk prediction. *Wirel. Internet Technol.* **2018**, *15*, 102–104.
25. Lei, X.M.; Xie, Y.T. Improved XGBoost Model Based on Genetic Algorithm for Hypertension Recipe Recognition. *Comput. Sci.* **2018**, *45*, 476–481.
26. Zhang, H.X.; Guo, H.; Wang, J.X. Research on Type 2 Diabetes Mellitus Precise Prediction Models Based on XGBoost Algorithm. *Chin. J. Lab. Diagn.* **2018**, *22*, 408–412.
27. Chen, T.; Guestrin, C. XGBoost: A Scalable Tree Boosting System. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data mining, San Francisco, CA, USA, 24–27 August 2016.