TELKOMNIKA Indonesian Journal of Electrical Engineering
Vol. 12, No. 9, September 2014, pp. 6909 ~ 6920
DOI: 10.11591/telkomnika.v12i9.5958

Improved SPRINT Algorithm and its Application in the Physical Data Analysis

Yazhi Ding[1], Zhigao Zheng[2], Rong Ma[1]*
[1]Institute of Physical Education, Xinjiang Normal University, Urumqi Xinjiang 830054, China
[2]School of Software and Microelectronics, Peking University, Beijing 102600, China
*Corresponding author, e-mail: xj_mr@hotmail.com (R. M)

Abstract

In order to determine the human physical condition according to the conventional tested data quickly and accurately, in this paper we proposed a trend selection based scalable parallelizable induction of decision trees algorithm (TESTSPRINT), based on the concept of pure interval and trend selection method. Based on the basic test data such as height, weight and grip strength, we can create a human physical condition decision tree quickly; according to the decision tree we can determine human physical health status quickly. Theoretic analysis and experimental demonstrations show that the algorithms this paper proposed outperforms existing algorithms in time and space complexity, and it was proved fruitful applications in the decision human physical health status with high accuracy.

Keywords*: SPRINT algorithm, gini index, physical data, data mining*

1. Introduction

Physical data analysis is providing a vital basis in aspects like developing national physical exercise and improving physical education. It has decided that it's necessary for us to offer decision-making support for other decision-related work in analyzing human physical data. In the past few years, the study in human physical data analysis is focused on the test method of physical data acquisition and instrument test, but stays general in analysis and statistical in comparing and verifying correlation for analyzing and processing on test data. Moreover, a lack of a deeper digging into data mining research and decision-making analysis on the enormous raw data collected in physical science study is making the deep implicit content hidden in test data more and more undiscoverable. Though statistics has achieved great obvious accomplishments in physical science research, it has revealed a lot of limitation on itself in the process of application data analysis which leads to the dissatisfaction in solving and analyzing large quantities of realistic test data. With the emergence of data mining technique, a scientific method of retrieving useful hidden information between data in enormous data set is discovered [1]. So far, the test on the index of judging human physical situation intuitively is relatively complex, let alone the high requirement on test method, then a new method is expected that we could estimate the physical state of subject simply just based on some simple test data in the actual normal physical state judgment. As a result, this new method has become an important topic in physical data analysis recently. The decision-making analysis has made a major breakthrough to this issue. Based on the decision-making method, a decision-making tree is built to directly determine the human physical data index based on some simple physical index like height, weight and so on, followed by the verification for the tree's accuracy with the use of partial test data. With the accurate tree, some simple measurement indexes are able to rapidly decide the human physical constitution status.

Yu Daifeng [2] and his fellows took the study based on the grip strength and muscle strength test data for example and applied the algorithm ID3 to the muscle strength data analysis which achieved a good effect in the stand-alone environment. But there exists the following shortages which remains much to be done in enormous data analysis and continuous data analysis [3].

1) Focused on the selection of features, prefer features with more characteristic value. But more value does not necessarily optimal, so it is not reasonable.

Received March 12, 2014; Revised June 27, 2014; Accepted July 15, 2014

2) Sensitive to noise and difficult to control the ratio between positive and negative examples.

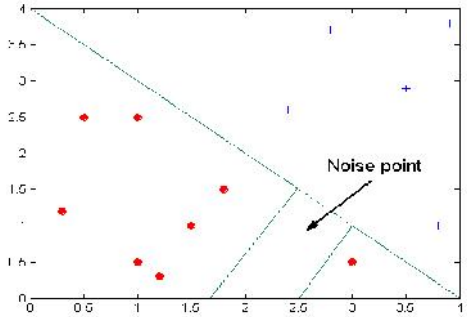

Figure 1. Influence of Noise Data for Decision Analysis

3) It is poor to learn simple logical expression and ID3 decision tree will change when the training set increases.

Because physical state monitoring data, like grip strength, blood pressure and BWH, are continuous data, the data in national data analysis is massive. Based on these disadvantages of ID3 algorithm and advantages of SPRINT algorithm in dealing with massive data analysis, this paper proposed trend selection based scalable parallelizable induction of decision trees algorithm, TESTSPRINT in short, to solve these shortages above. Theoretically and practically, the TESTSPRINT algorithm proposed here has the following advantages:

1) Kept the ability to support multi-CPU feature in SPRINT with low time and space complexity in dealing with massive data;

2) Introduced the concept of trend selection. Greatly decreased the computation complexity and online computation expense through pruning on attributes in continuous range by preprocessing based on trend selection theory.

3) Stayed low time and space expense while remaining high accuracy of physical data analysis.

This paper has established a solid theoretical foundation for discussing human physical situation by applying decision-making technique into human physical state analysis based on former data gathering. The second section firstly introduced the related work in aspects like human physical data analysis and SPRINT algorithm. The third part has a simple description of the principles and decision-making method in SPRINT algorithm. The fourth proposed TESTSPRINT algorithm based on trend selection. The fifth designed an experiment accordingly to obtain the physical quality standard by rapidly making decision analysis on some basic data like height, weight, grip strength, step index and so on. Then the accuracy of decision analysis can be measured by testing BMR value. Finally summary of this paper and prospects of future work are made.

2. Related Work

2.1. Physical Data Analysis

With the development of computer technique and deep-going of data mining technique, different scholars have adopted various data mining method to process physical status related data. Li [4] and his fellows analyzed the feasibility of data mining and data warehouse technique in analyzing high school physical data from the perspective of decision-making support from data warehouse. Sang [5] and his fellows applied data mining and logical organon in researching correlation between different management behaviors of instructors in Sports College and management effect on students and concluded that dynamically balanced management behaviors could improve the final effect of management work. Mao [6] and his fellows combined neutral network data mining and biochemistry index and analyzed the characteristics of neutral network self-learning to predict sports grades. Mao [7] and his fellows

combined gray ART aggregation analysis method theory and biochemistry index data to analyze and predict sports grades. The research provided a base for the quantization of the analysis, explanation and prediction on athletics stamina, improved the scientifically and intellectuality in training reserved athletes by coach while offering a scientific basis to adopt different scientific training program schema according to different athletics status of every athlete. Zhao [8] and his fellows put the model into analysis and assessment of tactics in sports competition, resulting in the successful prediction on the key factors in winning the competition. Zhen [9] and his fellows proposed a gastric cancer clinical medical information analysis application research model based on decision-making tree, utilized the SPRINT algorithm and constructed a dangerous factor analysis model of gastric cancer postoperative recurrence [10]. It revealed the chief dangerous factor of gastric cancer postoperative recurrence through analyzing the model and looking for the relationship between clinical diagnosis and prognostic.

2.2. SPRINT Algorithm

For the purpose of overcoming the disadvantages of SPRINT algorithm [11], different scholars listed a series of SPRINT variants from various perspectives which achieved wonderful effects. Liu [12] and his fellows proposed a numeric attributes processing method based on pure interval reduction to improve SPRINT. This method divides the range into multiple segments by using equal-width histogram, reduces between pure intervals and calculates precisely between non-pure intervals to guarantee splitting accuracy and decrease calculation amount. WU [13] and his fellows adopted similar method with Liu [12] which got a good result by applying the improved algorithm into graduation design process management system. Xu [14] and his fellows suggested a method to find the optimal break point rapidly to overcome the large calculation amount shortage in SPRINT accordingly. This method used a set of strategies like interval assessment, filtering and locality search to shrink the search space of SPRINT algorithm drastically. Peng [15] and his fellows improved the method of finding the optimal break point in continuous range to solve the large calculation amount in doing this by SPRINT algorithm. The improved one decreased the quantity of candidate splitting point, so the calculation amount and time got reduced. Yu [16] and his fellows introduced a dynamic data structure and applied the SPRINT algorithm in distributed environment, finally decreased the storage space for attribute list and the time consumed by splitting break point.

3. SPRINT Algorithm

3.1. Basic Idea of SPRINT Algorithm

The basic strategy of constructing decision-making tree is to adopt greedy method by top-down recursion and divide-and-conquer method. The construction of a decision-making tree consists of creation stage and pruning stage. The algorithm to construct the tree is as follows:

Algorithm 1:decision-making tree construction algorithm
Input: trained sample set T
Output: decision-making tree
Step 1: if T satisfies the extension condition, then returns;
Step 2: for every attribute A_i, find A_i's value or value set V_i, it will generate an optimal splitting point for test attribute A_i;
Step 3: compare every optimal splitting point of the attribute, choose the best one to divide T into T_1 and T_2;
Step 4: separately recur the decision-making tree of T_1 and T_2.

Figure 2. Decision Tree Building Algorithm

In the algorithm above, set T, T_1, T_2 each represents a node in the tree, among them T_1, T_2 are two child node of T. The final constructed tree a binary tree. SPRINT pruning adopted minimum description length principle [17].

3.2. Algorithm Implementation

Splitting index is a metric to measure the merit degree of attribute splitting rules. $Gini$ index is a kind of effective splitting index of searching optimal splitting point which is adopted in SPRINT algorithm. Suppose there is a data set S which holds n records, these records belong to c distinct classes, then the $Gini$ value of set S is:

$$Gini(S) = 1 - \sum_{j=1}^{c} p_j^2 \quad (1)$$

Which m is the number of records in set S, and S belongs to class j. Then $p_j = m/n$

If S is divided into two subsets S_1 and S_2 by the splitting rule $cond$, then the measurement value of this rule is marked as $Gini^D(S, cond)$, defined as Equation (2):

$$Gini^D(S, cond) = \frac{n_1}{n} Gini(S_1) + \frac{n_2}{n} Gini(S_2) \quad (2)$$

Where n_1 and n_2 each represents the number of records in S_1 and S_2. The smaller the value is, the better the splitting rule is.

For numeric attribute A, its splitting form is $A \le v$. As a result, these numeric attributes can be sorted firstly. Suppose that the sorted result is $v_1, v_2, \cdots, v_n$, because the splitting can only happen between two nodes, there exists $(n-1)$ possibilities. Usually the medium one $(v_i + v_{i+1})/2$ is chosen. Choose different splitting point from small to big successively and choose the one with minimum $Gini$ value as the optimal splitting point.

This method above can find the most accurate splitting point. But for these numeric attributes, the whole training set should be presorted, then choose the middle point value between every node as the splitting node to calculate the $Gini$ value. The temporary storage space consumed in splitting process is triple as the data storage space. The workload is so heavy that it leads to inefficiency for super large data set especially when there exists many values for each attribute.

4. SPRINT Algorithm Based on Trend Selection

Obviously the SPRINT algorithm has the shortage that the workload is heavy in accurately locating optimal splitting point. This paper introduces the pure interval concept in cited works [12], which splits the training continuous attributes into q segment by using equal-depth histogram [18], reduces between pure intervals, estimates the possibility of finding the optimal splitting point between non-pure intervals firstly and go on finding in most-possible segment. At last, the workload is decreased.

4.1. Basic Definition and Property

Definition 1. For a numeric attribute, if all records between range $[v_b, v_t]$ belong to the same class C_i, then this range is a pure range.

Theorem 1. If $f(x)$ is continuous in the range $[a,b]$, and there exists a first-order and second-order derivative in the range (a,b), then:

1) if $f''(x) > 0$ in the range (a,b),then $f(x)$ is concave in the range (a,b);
2) if $f''(x) < 0$ in the range (a,b),then $f(x)$ is convex in the range (a,b);

Theorem 2. $f(x)$ is a convex function in the pure interval.

Proof: suppose there exists a range $[v_l, v_u]$ in the remain-to-be-divided data set *S*, where:

n is the capacity of data set S;
c is the number of class in S;
x_i is the number of records whose value is less than or equal to v_l in class i;
y_i is the number of records whose value is less than or equal to v_u in class i;
C_i is the number of all records in class i;
n_l is the number of records whose value is less than and equal to v_l;
n_u is the number of records whose value is less than and equal to v_u;

According to equation (2), the value of $Gini^D$ value at point v_l is as follows:

$$Gini^D(S, a \le v_i) = \frac{n_l}{n}(1 - \sum_{i=1}^{c}(\frac{x_i}{n_l})^2) + \frac{n - n_l}{n}(1 - \sum_{i=1}^{c}(\frac{c_i - x_i}{n - n_l})^2) \quad (3)$$

$$\frac{\partial Gini^D(S, a \le v_i)}{\partial x_i} = \frac{2}{n_l(n - n_l)}(c_i \frac{n_l}{n} - x_i) - \frac{1}{n}(\frac{1}{(n - n_l)^2}1\sum_{i=1}^{c}(c_i - x_i)^2 - \frac{1}{n_l^2}\sum_{i=1}^{c}x_i^2) \quad (4)$$

For a pure range $[v_b, v_t]$ of C_k, in the equation (3), $x_i(i = 1, 2, \cdots, c)$ represents that the number of records in kth class is x_k, when only x_k changes, make $x_k = x, c_k = C$, then equation (3) can be transformed as a function of x. For Equation (3), make:

$$n_l = \sum_{i=1}^{c} x_i = A + x, \sum_{i=1}^{c} x_i^2 = B + x^2, \sum_{i=1}^{c}(c_i - cx_i)^2 = D + (c - x)^2$$

Where A, B, C, D are constant which over 0, then Equation (3) can be transformed as a function of x. Just as Equation (5) shows:

$$f(x) = 1 - \frac{B + x^2}{n(A + x)} - \frac{D + (C - x)^2}{n(n - A - x)} \quad (5)$$

The first-order derivation of Equation (5) is:

$$f'(x) = \frac{1}{n}(\frac{B + A^2}{(A + x)^2} - \frac{D + (n - A - x)^2}{(n - A - x)^2})$$

The second-order derivation of Equation (5) is:

$$f''(x) = -\frac{2}{n}(\frac{A^2 + B}{(A + x)^3} + \frac{(n - A - x)^3 + D}{(n - A - x)^3}) \quad (6)$$

As A, B, C, D all are constant values over or equal to zero, $x > 0, n > n_i = A + x, (A + x)^3$ and $(n - A - x)^3$ are both values over zero, then $f''(x) < 0$. According to theorem 1, it's known that the function $f(x)$ in the range $[v_b, v_t]$ are convex function.

Theorem 3. In the attribute table of continuous attribute, if the class of the two tuples which decide the splitting point v are the same, there exists at least another splitting point v' whose $Gini$ value is smaller than the $Gini$ value of splitting point v.

According to theorem 2, it is known that the minimal value in the pure interval $[v_b, v_t]$ of C_k can only be possible to exist at the border point of the interval. Numeric attributes are generally Gaussian distribution [19], so there are a lot of pure intervals divided in equal width interval. As a result, pruning can be done directly between pure intervals as long as the $Gini$ value of each pure interval's border point is calculated; for those non-pure intervals, firstly judge if the class of the two tuples which decide the splitting point v are the same, only when it is not, the calculation of the $Gini$ should be done.

4.2. Trend Selection Method

As the attribute value of $Gini^D$ is continuous, trend selection method should be firstly adopted to estimate the lower limiting value of $Gini^D$ for those non-pure intervals. Suppose for j, the minimal slope exists at point v_1, replace x_j with y_j, use the Equation (3) to calculate the new value of $Gini^D$. If $Gini^D(S, a \le v_1) > Gini^D(S, a \le v_u)$, then the value of $Gini^D$ in the range $[v_l, v_u]$ is convex which results in the pruning of this range. The value of $Gini^D$ candidate splitting point is $\min\{Gini^D(S, a \le v_1), Gini^D(S, a \le v_u)\}$. The trend selection based candidate segmentation algorithm is as follows:

Algorithm 2: TESTCASE

Input: The data set size n, number of records less than or equal to v_l is n_l, the number of intervals c.

Output: Candidate splitting point v and $Gini_{low}$ which is the value of $Gini^D$

Step 1 Initialize the candidate set $cs = \{1, 2, \cdots, c\}$

Step 2 Set $Gini_{low}$=1, calculate the $Gini^D$ of each interval's border point, choose the minimal value as $Gini_{can\_low}$

Step 3 Substitute every class in cs into equation (4), let the slope of j be minimal

Step 4 Replace x_j with y_j in vector x, then calculate $Gini^D$

Step 5 If the new $Gini^D$ is less than $Gini_{low}$, then replace $Gini_{low}$ with the new $Gini^D$, and delete j from cs, finally turn to step 3.

Step 6 Estimate $Gini_{low} = \min\{Gini_{low}, Gini^D(S, a \le v_l), Gini^D(S, a \le v_u)\}$

Step 7 If $Gini_{low} > Gini_{can\_low}$ then delete j, then to step 3.

Figure 3. TESTCASE Algorithm

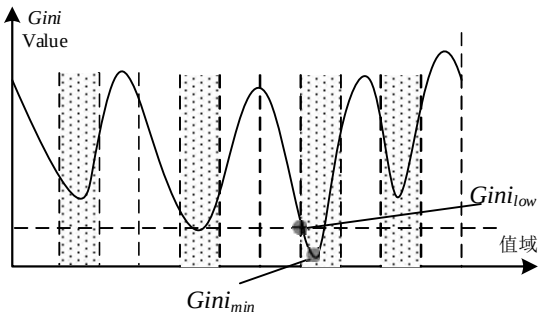

Figure 4. Candidate Interval and Estimate *Gini* Lower

At last, search the optimal splitting point in the candidate range. Figure 4 displays the range distribution and the situation $Gini_{low}$ and $Gini_{can\_low}$, in the example region, the columnar with shadow are candidate interval.

4.3. The SPRINT Algorithm Based on Trend Selection

Based on trend selection method, this paper proposes a SPRINT algorithm based on trend selection (TESTSPRINT in short). This algorithm adopts width first strategy to build decision tree and evaluate numeric attribute with the use of $Gini$ index. The proposed algorithm deals with numeric attribute through pure interval reduction method and trend based selection method, decreases computation amount with the help of pruning on the range of continuous attribute, and finally pick the optimal splitting point in a short time. Detail description of TESTSPRINT is as follows:

Algorithm 3:SPRINT algorithm based on trend selection
Input: Training set sample T
Output: Decision-making tree
Step 1 If T satisfies the stopping extension condition, then returns
Step 2 For those discrete attributes, scan the attribute list, refresh the count matrix to decide the optimal splitting subsets of these attributes
Step 3 for those continuous attributes, divide the attribute into q ranges using equal-width histogram, and build a partition histogram list
Step 4 Calculate the $Gini$ value of each pure interval's border point
Step 5 Prune on intervals using TESECASE algorithm
Step 6 Judge if the split tuples divided by the splitting point in the candidate interval are the same, if yes, delete it.
Step 7 Find the optimal splitting point of the whole attribute
Step 8 Compare every attribute's optimal splitting, choose the best one and divide T into T_1 and T_2
Step 9 Build decision-making tree for T_1 and T_2 recursively

Figure 5. TESTSPRINT Algorithm

In this algorithm, set T, T_1 and T_2 represents a node of the tree separately and T_1 and T_2 are the child node of T. Because there is only one optimal splitting point for each attribute, the final decision tree is a binary one.

4.4. Algorithm Analysis

As TESTSPRINT algorithm has done many pruning between intervals and on candidate splitting points, that is to say, the main advantages of this algorithm is constructing decision tree rapidly while processing splitting problems on continuous attributes reasonably. The main analysis on algorithm is focus on comparing the I/O requirement and time expense in building decision tree using two different algorithms.

Suppose there exists n records in data set S, these records belong to c distinct classes and are divided into q ranges in which the number of records is n_i, the number of non-pure intervals in the improved algorithm is b.

(1) Preprocessing stage

When the attribute table is constructed in the SPRINT algorithm, one read operation and one write operation are needed, its time complexity is $O(n)$; for each numeric attribute, two read operation and two write operation are executed in presorting at the expense of $O(n\log n)$.

In algorithm TESTSPRINT, attribute table can be built with one read and write operation, one more write operation is needed for constructing partition histogram table. Next preprocessing stage consumes $O(n)$ time.

(2) Node construction stage

In SPRINT algorithm, in order to find the optimal splitting point for a numeric attribute, a write operation should be done to the whole attribute table, so its time complexity is $O(n)$; it takes $O(n)$ time to divide the attribute table and this action consists of one write operation and one read operation.

In TESTSPRINT algorithm, estimating the $Gini$ value of each interval's border point takes $O(qc)$ time; it needs $O(n)$ time to decide every non-pure interval's records and to construct a temporal interval attribute table; the worst case for calculating accurate $Gini$ value takes $O(\sum_{i=1}^{b} n_i \log n_i + n_i c)$; $O(n)$ time is consumed while splitting the attribute table.

In SPRINT algorithm, the main time expense is used to sort all records of the attribute table in the whole process. The TESTSPRINT avoids overall sorting effectively and only go on local sorting for non-pure intervals. At the same time, reduction on pure intervals, partial pruning on non-pure intervals and pruning on candidate nodes reduce the calculation amount of $Gini$ value.

5. Experiment Verification

In this section, we will compare other algorithms to research the performance and effectiveness of the SPRINT algorithm based on trend selection through experiments. In this experiment, the data used is made up of the physical measurement data of 225 subjects in Xinjiang Normal University and the usually used STATLOG data set for decision support. The data items include height, weight, grip, step index and BMR value in which training set involves 150 tuples and test set includes 75 tuples. In the STATLOG data set, the Segment data set includes 2310 records, Shuttle includes 58000 records and Satimage includes 6435 records.

All experiments in this paper are done in the same hardware and software environment. The specific environment is Intel(R) Core(TM) i5-25200 quad-core 64 bits 2.5GHz CPU and 4GB memory. The software environment is Windows 7-64 bits(professional) OS, all code is written in Java(64 bit JDK) and Matlab (2012).

5.1. Algorithm Verification

To verify the accuracy and stability of the algorithm, this paper takes advantage of STATLOG to testify the algorithm's accuracy. The result is in Table 1.

Table 1. Exact $Gini$ Value and Compute Value under Different Intervals of TESTSPRINT

Data set	Accuracy value	Number of intervals				
		200	100	20	25	10
Segment	0.714286	0.714286	0.714286	0.714286	0.714286	0.715799
Shuttle	0.175777	0.175777	0.175777	0.175777	0.175777	0.175777
Satimage	0.653167	0.653167	0.653167	0.653167	0.653167	0.653167

Just as Table 1, the splitting result of TESTSPRINT and SPRINT are basically the same. The accuracy of TESTSPRINT reaches 99% or more, the result of accuracy is acceptable.

5.2. Data Acquisition

In the test, we use HHIC/WL-100 Grip Measurement Tester to measure the power of grip of the subjects. Power Scope: 5kgf ~99.9kgf, measurement distinguishability is 0.1kgf, measurement error: ±0.3kgf. The subject holds the tester tightly, rotate grip distance adjusting

knob to form a 90 degree by bending the second joint of index finger. Separate two feet naturally, stay orthostatism, hang down two limbs, palm inward, don't move arms from side to side, no touch between tester and any part of the subject's body, hold the tester with all his strength. Two testsare done for each subject and the maximal record is chosen.

HHTC/TJ-100 step instrument is used to collect step index. Measurement scope: 0-300 times/minute, measurement distinguishability: one time, measurement error: ±1time. The subject can do some slight preparation activity before tested, mainly for lower limbs. The subjects do an up-down step every two seconds with the pace of the music. The pace of the music is 120times/minute, 4 pace/time, the duration lasts for 3 minutes. The step height for boys is 40cm and 35cm for girls. After the movement gets stopped, the subject sit quietly for 30 seconds, put the palms on the desktop, keep the fingers at the same height of the heart as possible as the subject can. The conner tests the pulse of the subject using pulse testing instrument. The test time lasts for 3.5 minutes until the test result is recorded.

The BMR measurement is tested by using InBody3.0 body composition analyzer in basic condition.InBody3.0 takes no more than 2 minutes. Weight measurement scope: 10kg~25kg, age scope: 6~99 years old. InBody3.0 body composition analyzer test project consists of BMR, BMI and weight (kg).

The above test results can be summarized to obtain the subjects' corporeity. The format of test data is just as Table 2 displays.

Table 2. Example Data Format

Age	Sex	Height (cm)	Weight (kg)	Grip (kgf)	Step index (time/minute)	corporeity	BMR (J/(h·m^2))
21	F	150	59.2	16	47	1	1300
20	F	170	55.2	27.8	52	2	1548
21	F	170	59.7	21.7	42	2	1501
19	F	163	53.9	24.1	52	2	1477
21	M	170	56.7	40.3	51	2	1604
19	M	175	54	37.2	57	1	1628
22	M	170	62.6	59.4	81	3	1793
20	F	162	58.5	35.7	46	3	1598
20	M	173	67.5	39.8	63	3	1952

5.3. Experiment Analysis

In ordinary body status judgment, we hope we can decide the body status of the subject according to some simple measurement data. Thought BMR can achieve this purpose clinically, it is complex at a high cost, making it not suitable for rapid test. As a result, we designed this experiment according to the TESTSPRINT algorithm which takes height, weight, grip, step index and some other basic data into consideration to support rapidly making decision on body status. Finally, the accuracy of the result can be obtained by testing BMR values in training set.

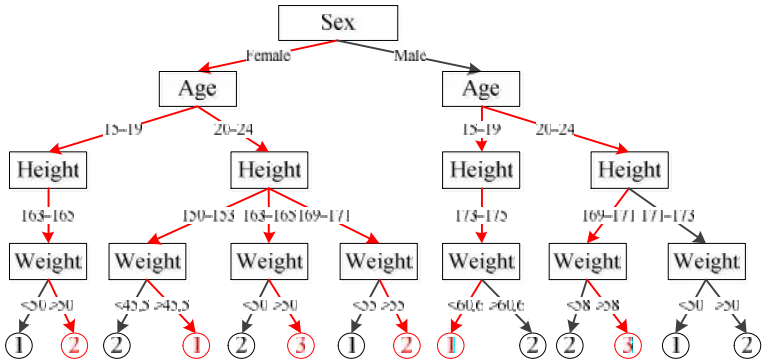

Figure 6. Decision Tree Build by the Original SPRINT Algorithm

There exists five description types like excellent, fine, healthy, sub-health and unhealthy, then we use 0 to 4 to represent each type separately. Consequentially, we divided the data set into five intervals to calculate the result in the experiment. The decision tree constructed by raw SPRINT through example data is shown in Figure 6.

At the beginning, simulated analysis is done on training set data and test set data separately. The results are displayed in Figure 7 and Figure 8 separately.

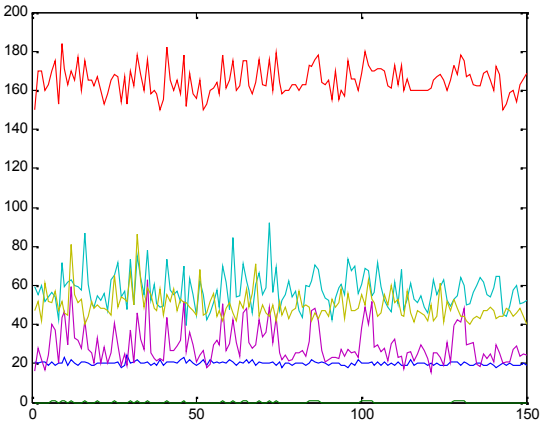

Figure 7. Training Set Data Diagram

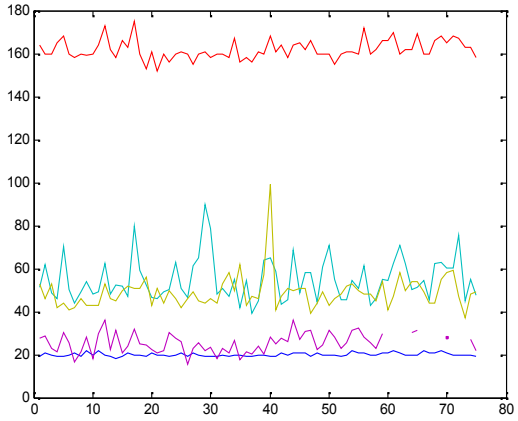

Figure 8. Test Set Data Diagram

From the start, we model the decision tree based on training set data, the estimation on test set data is done later. Three categories and one outlier data are discovered. The accuracy of the algorithm reaches 99.56% and the accuracy of raw SPRINT reaches 98.2%.

Wherein, the raw SPRINT and the proposed TESTSPRINT have different residual with its residual interval graph which is shown in Figure 9 and Figure 10.

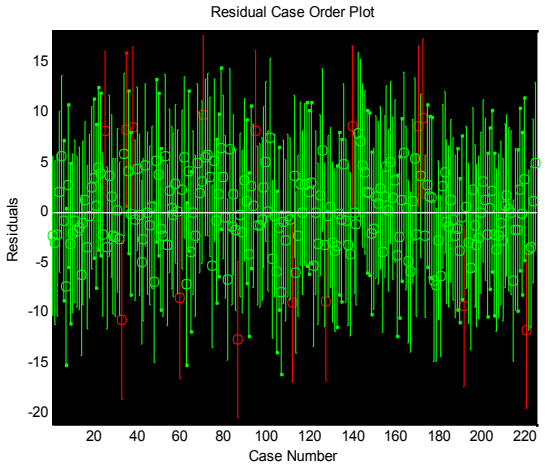

Figure 9. Residuals and Residual Interval Graph of Original SPRINT Algorithm

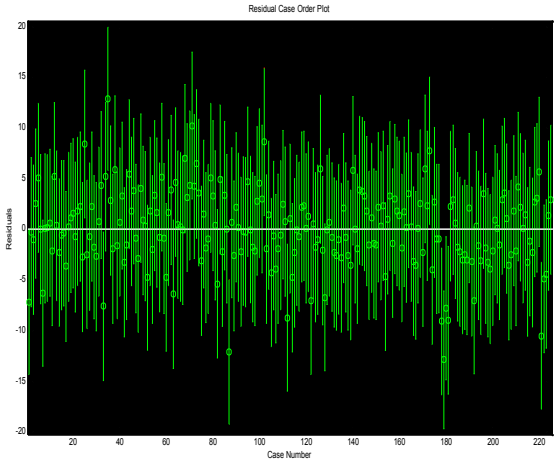

Figure 10. Residuals and Residual Interval Graph of TESTSPRINT Algorithm

It is known in Figure 9 and Figure 10 that result from raw SPRINT has some noise data which influences the accuracy of the result, but an improved result is obtained from TESTSPRINT algorithm proposed in this paper. A conclusion is made that the TESTSPRINT algorithm has a high accuracy in analyzing physical status data and its prediction ability is acceptable.

6. Conclusion and Future Work Prospects

6.1. Conclusion

This paper introduced the concept of pure interval, put forward trend selection method and proposed the TESTSPRINT algorithm. There exists the following characteristic: G_{ini} index function in the pure intervals is convex, numeric attribute follow the rule of Gaussian distribution [20] and first-order derivative of G_{ini} exponential function can predict trend. Something is done to optimize the numeric attribute in SPRINT algorithm with the help of above facts. It is proved in the algorithm analysis that TESTSPRINT is an effective method. Of course, the numbers of divided intervals do have some impact on the time consumed by the algorithm. So more has to be done in picking proper size of intervals effectively to decrease the number of non-pure intervals and sorting expense between intervals. What's more, some optimization can be done on pruning the candidate splitting point of non-pure intervals.

6.2. Future Work Prospects

With more and more importance attached to national physical quality by government and citizens, physical monitoring data system tends to share data between different regions. So far, our country has begun to build national physical and geographical information system [21], badminton players' tactical awareness test and other multi-media training systems [22]. In the future, as the data size in DBS increases, the analysis for national physical data will adopt distributed big data computation. As a result, more research should be done on distributed big data computation while optimizing local data processing. On the other side, simple statistics is used in current physical data analysis. For the purpose of mining deep connotation of physical data to obtain more accurate conclusion, much more should be done in data mining algorithm and its application in physical data analysis.

Acknowledgements

The research projects of the Humanity and Social Science Youth Foundation of Ministry of Education of China No.11XJJC840001, and The Natural Science Foundation of Jiangsu Province under Grant No.BK2010139.

References

[1] Boris Milovic, Sava Kovacevic jsc Vrbas. Prediction and decision making in Health Care using Data Mining. *International Journal of Public Health Science (IJPHS).* 2012; 1(2): 69-78
[2] Yu Daifeng, Zhong Yaping,Yu Yaguang. Application in body muscle strength data analysis based on data mining technique--take body grip muscle strength test data research for example. *Sports Science.* 2010; 30(2): 70-74.
[3] JR Quinlan. C4.5: Programs for machine leaning. Los Altos, California: Morgan Kaufmann Publishers, Inc. 1993.
[4] Li Huiling,Lin Zi. The application of data warehouse and data mining in high school physical data analysis. *Journal of Guangzhou Institute of Physical Education.* 2005; 25(5): 126-128.
[5] Sang Guoqiang. Fuzzy analysis of instructors' management behavior effect in institute of physical education. *Journal of Wuhan Institute of Physical Education*. 2013; 47(2): 81-86.
[6] Mao Jie, Jiang Xiongwen, MeiYan. The application of neutral network in sports biochemistry index. *Wuhan sports academical journal*. 2004; 38(4): 53-55.
[7] Mao Jie, MeiYan. The application of gray ART clustering methodology in competitive sports biochemistry index monitoring. *Journal of Wuhan Institute of Physical Education*. 2005; 39(10): 50-52.
[8] Zhao Huiqun, Sun Jing, Hua Yongmin,Jin Jichun. The applied research of data mining technique in sports contest's tactical analysis. *Journal of Beijing Sports University.* 2008; 31(5): 712-715.
[9] Zhen Danqing. Gastric cancer clinical medical data mining research based on SPRINT algorithm. *Journal of Jilin normal university (nature and science edition).* 2012; 33(2): 121-124.
[10] Mehta M, Agrawal R, Rissanen J. *SLIQ: A fast scalable classifier for data mining*[C].In: Proceedings of 1996 International Conference on Extending Databases Technology, Avignon, France. 1996: 18~32.
[11] Shafer J, Agrawal R, Mehta M. SPRINT: A scalable parallel classifier for data mining[C].In: Proceedings of the 1996 International Conference on Very Large Data Bases. Bombay, India. 1996: 544~555.
[12] Liu Youjun, Wang Linlin. Improvement in SPRINT algorithm. *Computer Engineering*. 2006; 32(16): 55-57.

[13] Yan-Wen Wu, Li Li, Sheng-Yi Zhao, Xue-Yi Ai. *Application of Improved SPRINT Algorithm in the Graduation Design Process Management System*. Workshop on Intelligent Information Technology Application, Zhang Jiajie, China. IEEE computer society. 2007; 252-255.
[14] Xu Xiangyang, Gong Yonghua. Improvement In SPRINT algorithm. *Computer Engineering and Application*. 2003; 39(33): 187-189.
[15] Peng Cheng, Luo Ke. Improved SPRINT algorithm in finding splitting point of continuous attribute. *Computer Engineering and application.* 2006; 42(27): 159-161.
[16] Yu Lei, Liu Dayou, Gao Ying, Tian Ye. Improved SPRINT algorithm and its application in distributed environment. *Journal of Jilin University(Natural Science edition)*. 2008; 46(6): 1119-1124.
[17] Gerardo Richarte. Four different tricks to bypass Stack Shield and Stack Guard protection. http://www2.corest.com/files/files/11/StackguardPaper.pdf. 2002.
[18] U Fayyad, K Irani. *Multi-interval Discretization of Continuous-values attributes for Classification Learning.* Proc 13th Intl Joint Conf on Artificial Intelligence, Chambery, France. 1993.
[19] Han J, Kamber M. Data Mining: Concepts and Techniques. Beijing: High Education Press, 2001: 279-301.
[20] Indah Agustien Siradjuddin, M. Rahmat Widyanto, T. Basaruddin. Particle Filter with Gaussian Weighting for Human Tracking. *TELKOMNIKA Indonesian Journal of Electrical Engineering*. 2012; 6(10): 1453-1457.
[21] Shi Bin, Wang Yuhong. The initial research on constructing national sport geography information system. *Journal of Xi'an sport college*. 2007; 24(2): 1-8.
[22] Cheng Yongmin, Jin Hua, Zhou Weixing, Hu Xiaohui. Measurement of badminton players' tactical awareness and the research on multimedia training system. *Journal of Guangzhou sports college*. 2009; 29(2): 57-61.