Research Journal of Applied Sciences, Engineering and Technology 11(9): 1019-1023, 2015
DOI: 10.19026/rjaset.11.2143
ISSN: 2040-7459; e-ISSN: 2040-7467

Submitted: June 24, 2015 Accepted: August 15, 2015 Published: November 25, 2015

Research Article

Comparison of Two New Data Mining Approach with Existing Approaches

[1]Jun Zhang, [1]Junjun Liu and [2]Qing E. Wu
[1]School of Information Engineering, Zhengzhou University of Science and Technology, Zhengzhou, 450064, China
[2]College of Electric and Information Engineering, Zhengzhou University of Light Industry, Zhengzhou, 450002, China

Abstract: This study studies two uncertainty data mining approaches and gives the two algorithms implementation in the software system fault diagnosis. We discuss the application comparison of the two data mining approaches with four classical data mining approaches in software system fault diagnosis. We measure the performance of each approach from the sensitivity, specificity, accuracy rate and run-time and choose an optimum approach from several approaches to do comparative study. On the data of 1080 samples, the test results show that the sensitivity of the fuzzy incomplete approach is or so 95.0%, the specificity is or so 94.32%, the accuracy is or so 94.54%, the run-time is 0.41 sec. Synthesizing all the performance measures, the performance of the fuzzy incomplete approach is best, followed by decision tree and support vector machine is better and then followed by Logistic regression, statistical approach and the neural networks in turn. These researches in this study offer a new thinking approach and a suitable choice on data mining.

Keywords: Data mining approach, fuzzy incompleteness, performance indexes, statistical approach

INTRODUCTION

Because of the rapid increase of measurement data in engineering application and the participation of human, the uncertainty of information in data is more prominent and the relationship among data is more complex. How to mine some potential and useful information from plentiful, fuzzy, disorderly and unsystematic, strong interferential data, so as to perform real-time and effective engineering applications, this is a problem needs to be urgently further study.

Data mining is a process of selection, exploration and modeling to a mass of data for discovering beforehand unknown rules and relations, whose purpose is to get some clear and useful results for the owner of the database presented by Giudici *et al*. (2004). The spread speed of data mining was very fast and its application scope was widespread day by day introduced by Giudici *et al*. (2004), Liang (2006), Zhang *et al*. (2008) and Chen *et al*. (2008). Liang (2006) provided several data mining algorithms and some applications in engineering. Zhang *et al*. (2008) and Chen *et al*. (2008) introduced three data mining algorithms in medicine applications. However, the data mining industry was still in the initial stage of development in China, the domestic industries basically didn't have their own data mining systems.

Now, some algorithms on data mining have been relatively mature as shown in Balzano and Del Sorbo (2007) and Wolff *et al*. (2009). The decision Tree algorithm based on CHAID, some rules generated by Scenario could be applied to the unclassified data set to predict which records would have promising results. Scenario's decision tree algorithm is very flexible, which gives the user the choice to split any variable, or the choice of splitting with statistical significance. He carried out the graphical analysis to the crude data by using the fold line chart, histogram and scatter plot. Liang (2006) listed several main software developers on data mining.

This study introduces two new approaches on data mining, uses them and other classical supervised learning data mining technologies to learn and classify 1080 data, validates the feasibility and effectiveness for the new data mining approach and compares the performance of these approaches with each other, so as to hope that can select a best mining approach for fault diagnosis in software system. This study evaluates the performance of each approach from sensitivity, specificity, accuracy, respectively.

FUZZY INCOMPLETE APPROACH

In here, the positive region and the reduction are mainly used. Their definitions refer to Pawlak (1982).

Corresponding Author: Qing E. Wu, College of Electric and Information Engineering, Zhengzhou University of Light Industry, Zhengzhou, 450002, China

The fuzzy incomplete approach consists of three procedures and is given as follows.

Firstly, the incomplete reduction algorithm is as follows:

Input: a set of condition attribute is C = {a_1, a_2,…, a_n} and a set of decision attribute is D = {d}.

Output: A set of attribute reduction is *RED* (Ω).

Step 1: Compute the C positive region of D is $POS_C(D)$.

Step 2: For an attribute $a_i \in C$, after it is removed, the obtained subset of condition attribute is C\ {a_i}. Then compute the C\ {a_i} positive region of D is $POS_{(C\setminus\{a_i\})}(D)$.

Step 3: If $POS_{(C\setminus\{a_i\})}(D) = POS_C(D)$, then it indicates the attribute a_i to relative to the decision attribute D is unnecessary. Assign C = C\ {a_i} and go to the step 2. Otherwise, the attribute reduction $RED(\Omega) = C$ is output.

Secondly, define α_i (k) is the deviation between the measured attribute and the necessary attribute, for example, it is a norm or a covariance of error.

If we choose the normal membership function, the similar degree of the deviation under the normal operating condition at the moment k is:

$$d_i(k) = e^{-b\alpha_i(k)}$$

where, the $0 < b \le 1$ is a pending constant. Obviously there is $0 \le d_i(k) \le 1$.

When $\forall k \in \{1,2,\cdots,l\}$, after the $d_i(k)$ is obtained, the similarity vector between the standard value of the necessary attribute and the measured value of the real state in the normal operating condition is also obtained and labeled as M_i (l), i.e.:

$$M_i(l) = \left(d_i(1), d_i(2), \cdots, d_i(l)\right)'$$

where $M_i(l) \in [0,1]^l$.

Based on the definition of fuzzy synthetic function, define the synthetic similar degree of the deviation from time 1 to time l is:

$$\beta_i(l) \triangleq S_l\left(M_i(l)\right) = S_l\left(\left(d_i(1), d_i(2), \cdots, d_i(l)\right)'\right)$$

where,

$$S_l\left(M_i(l)\right) = \left(\frac{1}{l}\sum_{k=1}^{l} d_i^q(k)\right)^{\frac{1}{q}},\ q > 0$$

Thirdly, in order to give the similarity judgment between the standard value and the measured value, we assume H_0 and H_1 are the following event:

H_0: If the similar degree is bigger than a certain threshold value, a certain attribute is a necessary attribute.

H_1: If the similar degree is not bigger than a certain threshold value, a certain attribute is an unnecessary attribute.

Assume the threshold parameter is ξ. According to the experience and the test, there is $0.5 \le \xi \le 1$. If:

$$\beta_i(l) > \xi$$

Then H_0 is accepted, i.e., a certain attribute is a necessary attribute; otherwise H_1 is accepted. Therefore, some data of the test similarity that satisfy the above formula are required. Otherwise some unsatisfactory data are removed.

MINING OF UNKNOWN PARAMETERS CALLED A STATISTICAL APPROACH

This study implements the data mining to the unknown parameters by the characteristics of statistical distribution. Because many random variables in practice problems obey (or approximately obey) a normal distribution, this study focuses on the introduction of the mining of unknown parameters about the normal population.

Let $X_1, X_2, \cdots, X_n$ be a sample with n capacities from a normal population N (a, σ^2).

Here give an instance whether a mean is equal to the mining of known value.

This mining problem is:

$$H_0 : a = a_0$$

$$H_1 : a \ne a_0$$

Here, a_0 is a known mean. The σ^2 is an unknown. The σ^2 is discussed as follows:

When the H_0 comes into existence, $\frac{\bar{X}-a_0}{\sigma/\sqrt{n}}$ contains an unknown parameter σ, but $S^{*2} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - X)2 = nn-1S2$ approximates to σ^2 better, so we think naturally that we may use the $S^* = \sqrt{S^{*2}} = \sqrt{\frac{n}{n-1}}S$ to replace the parameter σ in $\frac{\bar{X}-a_0}{\sigma/\sqrt{n}}$. Thus, the statistic $\frac{\bar{X}-a_0}{\sqrt{\frac{n}{n-1}}S/\sqrt{n}} = \frac{\bar{X}-a_0}{S} \bullet \sqrt{n-1}$ is obtained. According to

Table 1: The confusion matrix of performance index

	Actually belong to the correct	Actually belong to the error
Classified as the correct samples	N_{c1}	N_e
Classified as the error samples	N_m	N_{c2}

the statistical theory, when the H_0 comes into existence, there is $\frac{\bar{X}-a_0}{S}\sqrt{n-1}\ T \sim t(n-1)$. So, for a given significant level $\alpha(0<\alpha<1)$, the critical value t_a of the distribution with the freedom degree n-1 can be obtained by the t-distribution table so as to make $P(|T|>t_\alpha)=\alpha$.

For a given sample observed value x_1, …, x_n, we calculate the value of $T=\frac{\bar{X}-a_0}{S}\sqrt{n-1}$ is $t=\frac{\bar{x}-a_0}{s}\sqrt{n-1}$. The mining method is:

If $|t|>t_\alpha$, then H_0 is rejected, otherwise H_0 is accepted. The mining method is called the T-mining method.

COMPARISON OF THE NEW AND SEVERAL EXISTING DATA MINING APPROACHES

Criteria of performance index: The confusion matrix is used for calculating the classification accuracy. To the classification of 2 categories as an example, the confusion matrix is shown in Table 1.
That is, the confusion matrix is:

$$C_M=\begin{pmatrix} N_{c1} & N_e \\ N_m & N_{c2} \end{pmatrix}$$

Table 1, the N_{c1} is the number of samples for the first kind correct classification, i.e., denotes the number of the samples that actually belong to the correct and are also classified as the correct. The N_m is the number of samples for the missed classification, i.e., denotes the number of the samples that actually belong to the correct, but classified as the error. The N_{c2} is the number of samples for the second kind correct classification, i.e., denotes the number of the samples that actually belong to the error and are also classified as the error. The N_e is the number of samples for the error classification, i.e., denotes the number of the samples that actually belong to the error but classified as the correct.
The sensitivity is calculated by:

$$N_1=\frac{N_{c1}}{N_{c1}+N_m}$$

The specificity is calculated by:

$$N_2=\frac{N_{c2}}{N_{c2}+N_e}$$

Moreover, assume $N=N_{c1}+N_m+N_{c2}+N_e$ and $N_c=N_{c1}+N_{c2}$. The correct classified rate, the error classified rate and the missed classified rate are defined as follows:

$$P_c=\frac{N_c}{N},\ P_e=\frac{N_e}{N},\ P_m=\frac{N_m}{N}$$

where the N denotes the total number of samples. The N_c is the number of samples for the correct classification. The P_c denotes the correct classified rate. The P_e denotes the error classified rate. The P_m denotes the missed classified rate. Then, obviously, there exists

Table 2: Test results of performance of fuzzy incomplete and statistical approaches

Group No.	Total samples sample No.	Fuzzy incomplete Confusion matrix		Sensitivity	Specificity	Accuracy	Statistical approach Confusion matrix		Sensitivity	Specificity	Accuracy
1	192	112	4	0.9180	0.9429	0.9271	103	15	0.8512	0.7887	0.8281
		10	66				18	56			
2	296	190	7	0.9360	0.9247	0.9324	175	22	0.8578	0.7609	0.8277
		13	86				29	70			
3	396	247	8	0.9114	0.9360	0.9192	223	29	0.8383	0.7769	0.8182
		24	117				43	101			
4	499	339	10	0.9313	0.9259	0.9299	268	36	0.8701	0.8115	0.8477
		25	125				40	155			
5	526	334	12	0.9252	0.9273	0.9259	310	29	0.8158	0.8141	0.8308
		27	153				70	127			
6	607	372	12	0.9007	0.9381	0.9127	345	26	0.8175	0.8595	0.8303
		41	182				77	159			
7	712	457	20	0.9327	0.9099	0.9256	422	47	0.8810	0.7983	0.8539
		33	202				57	186			
8	813	507	19	0.9286	0.9288	0.9287	470	42	0.8672	0.8450	0.8598
		39	248				72	229			
9	929	592	17	0.9150	0.9397	0.9225	555	33	0.8486	0.8800	0.8579
		55	265				99	242			
10	1080	323	42	0.9500	0.9432	0.9454	291	75	0.9037	0.9011	0.9019
		17	698				31	683			
Mean				0.9249	0.9317	0.9269			0.8551	0.8236	0.8456
S.D.				0.0161	0.0092	0.0088			0.0281	0.0487	0.0282

Table 3: The performance indexes of each approach

Performance indexes	Fuzzy incomplete		Statistic		Neural networks		Support vector		Decision tree		Logistic regress	
Confusion	323	42	291	75	271	45	272	39	313	57	283	46
matrix	17	698	31	683	47	717	42	727	29	681	40	711
Sensitivity	95.00%		90.37%		85.22%		86.62%		91.52%		87.62%	
Specificity	94.32%		90.11%		94.09%		94.91%		92.28%		93.92%	
Accuracy	94.54%		90.19%		91.48%		92.50%		92.04%		92.04%	
Error rate	3.89%		6.94%		4.17%		3.61%		5.28%		4.26%	
Missed rate	1.57%		2.87%		4.35%		3.89%		2.69%		3.70%	
Runtime/sec	0.41		0.58%		1. 07		2.45		2.36		0.99	

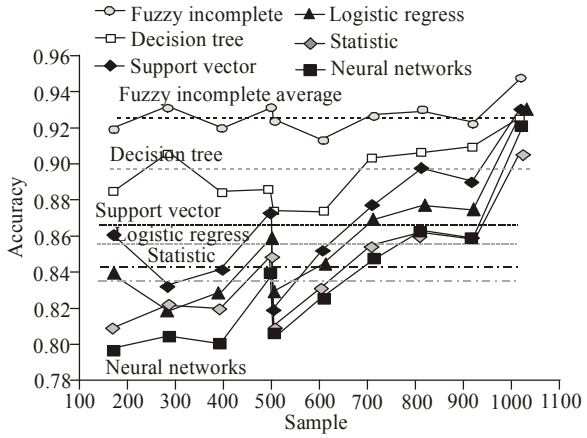

Fig. 1: The accuracy and the mean of 6 approaches

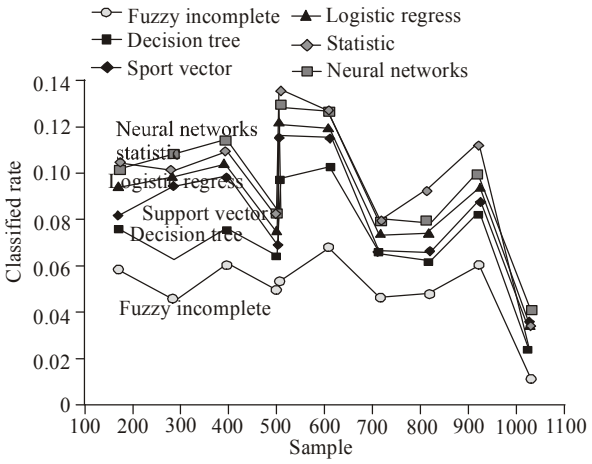

Fig. 2: The missed classified rate of 6 approaches

$P_c + P_e + P_m = 1$. We usually use their relative frequency instead of their probability in simulation. The above parameters all can reflect the performance of data mining scheme.

Experiment and comparison: Chen *et al*. (2008), Aburrous *et al*. (2010) and Khalifelu and Gharehchopogh (2012) gave the forecast accuracy of the decision tree approach was higher than the corresponding value of other approaches and its standard deviation was less than that of other approaches. But by experimental validation, these performances of the fuzzy incomplete approach introduced in this study are better than those of the decision tree approach. The conclusion is shown in the following.

The test results of two approaches on data mining are given in experiment in here, i.e., the test results of performance of the fuzzy incomplete and statistical approaches for 10-group samples of historical measured data of software aging, which are shown in Table 2. Similarly, the test results of other approaches can also be given. Here they are omitted.

We experiment with 10-group data to compare the two new approaches with the existing approaches given by Liang (2006), Chen *et al*. (2008), Aburrous *et al*. (2010) and Khalifelu and Gharehchopogh (2012), but for simplicity, the test results of only one sample set here are given, as shown in Table 3.

By the experiment of 10-group data, the results of correct performance to every approach are shown in Fig. 1 and 2.

In this study, we use six indexes which are sensitivity, specificity, forecast accuracy, error classified rate, missed classified rate and runtime to compare the performances of six data mining approaches. From Fig. 1 and 2 and Table 3 known, the fuzzy incomplete approach has the highest sensitivity, the forecast accuracy for every group, which is higher than those of other approaches. The average forecast accuracy of fuzzy incomplete approach is also slightly higher than that of the other approaches and its runtime is least. Moreover, in the test of small sample set, the standard deviation of the forecast accuracy, sensitivity and specificity of fuzzy incomplete approach in the 10 groups, mean of error classified rate and missed classified rate all are less than those of the other approaches, it indicates its forecast results are relatively stable. Therefore, the performance of the forecast model established by the fuzzy incomplete approach is slightly better than that of other models on the whole, so the fuzzy incomplete approach is a preferred approach. Secondly, the decision tree, support vector machine, logistic regression, statistical approach and neural networks are followed in turn.

CONCLUSION

This study uses the six data mining approaches to test 10-group data whose number of samples is 192,

296, 526, 929, 1080 and so on, respectively. A best performance is selected from every approach based on the sensitivity, specificity, accuracy, error classified rate, missed classified rate and running time to compare with each other, in order to discover a suitable approach for aging characteristic research of software system. The test results show that the fuzzy incomplete approach is the best, the next is the decision tree, followed by the support vector machine, Logistic regression and statistical approach, the worst is the neural network. Through the contrast research discovered, the fuzzy incomplete approach is more suitable for the research of characteristic discrimination of aging detection in software system.

ACKNOWLEDGMENT

This study is supported by Project of Henan Province Science and Technology (No: 142300410247); Key Project of Henan Province Education Department (No: 14A413002, 12A520049); Project of Zhengzhou Science and Technology (No: 131PPTGG411-4), respectively.

REFERENCES

Aburrous, M., M.A. Hossain, K. Dahal and F. Thabtah, 2010. Intelligent phishing detection system for e-banking using fuzzy data mining. Expert Syst. Appl., 37(12): 7913-7921.

Balzano, W. and M.R. Del Sorbo, 2007. Genomic comparison using data mining techniques based on a possibilistic fuzzy sets model. BioSystems, 88: 343-349.

Chen, J.X., G.C. Xi and W. Wang, 2008. A comparison study of data mining algorithms in coronary heart disease clinical application. Beijing Biomed. Eng., 27(3): 249-252.

Giudici, P., Y. Fang, W. Yu, W. Lijuan *et al*., 2004. Applied Data Mining Statistical Methods for Business and Industry. Electronics Industry Press, Beijing.

Khalifelu, Z.A. and F.S. Gharehchopogh, 2012. Comparison and evaluation of data mining techniques with algorithmic models in software cost estimation. Proc. Technol., 1: 65-71.

Liang, X., 2006. Data Mining Algorithm and its Application. Beijing University Press, Beijing.

Pawlak, Z., 1982. Rough sets. Int. J. Comput. Inf. Sci., 11: 341-356.

Wolff, R., K. Bhaduri and H. Kargupta, 2009. A generic local algorithm for mining data streams in large distributed systems. IEEE T. Knowl. Data En., 21(4): 465-478.

Zhang, L., Z. Gong, Y. Chen and S. Gu, 2008. Biomedical Data Mining. Shanghai Science and Technology Press, Shanghai, China.