Using Augmented Bayesian Networks to Compare Preference of Performance

Yong-Gyu Jung and Young-Jin Choi*

Eulji University, Department of {Medical IT Marketing, Healthcare Management}
212 Yangji Sujung Sungnam 461-713 Korea
{ygjung, yuzin}@eulji.ac.kr
**corresponding Author*

Abstract

The data mining technique is applied in various fields as a method to extract information based on massive data, and Bayesian networks are also utilized as useful modeling technique. Accordingly, many algorithms in Bayesian networks such as K2, TAN in expansion have been proposed, and suitability of algorithm for each situation evaluation stage has been requested based on performance test result validation to selectively use optimum algorithm for certain situation. As massive various that affects the result exists in actual situation, acquired information through certain data mining technique is considerably limited. Also, the filmed medical images may positively affect the diagnosis but due to high weight on subjective judgment, it is an abstruse problem to process with automatic system. Through this, improved expansion model of search algorithm is proposed with the K2 or TAN in Bayesian networks, which is relatively advantageous in handling the complicated situation of reality and is based on multivariate probability model. Now, because of the nature of extended Bayesian network which greatly varies the performance depending on the type of applied search algorithm, realistic evaluation is required on performance and suitability of each techniques. So in this thesis, experimentation by using equivalent data on disease diagnosis in extended Bayesian network is conducted, and measured classification accuracy while giving changes in search algorithm such as K2 and TAN. In the experiment, comparative evaluation of performance is done based on the result analysis of 10-fold cross validation, and made it possible to distinguish high risk data through classifying HRCT images of patients with high risk of reoccurring of the disease.

Keywords: *PCA, Random Projections, Data Conversion, Extended Bayesian Network, HRCT, K2, TAN*

1. Introduction

Recently, various industries own product characteristics, production history, preference research, and researches to make practical use of it is done. Data mining field is drawing the attention as it analyze the relations and characteristics of each attributes based on these numerous data, and model into information.

Various algorithms such as association, clustering, decision tree, neural network are researched in data mining field to extract information through statistic analysis or modeling of data [1]. However if these techniques are applied in actual problems, it is not a simple problem to select the optimum algorithm that works perfectly in all situations, due to existence of numerous factors that can affect results. Especially, researches on Bayesian network as a method to express knowledge based on probability is active recently. All phenomenons in actual situations results from the relations between attributes are closer to probability. As Bayesian network predicts the result by using probability on each node, result of prediction varies greatly depending on choices among search algorithms such as K2 or TAN (Tree Augmented Naïve Bayes). By applying K2 and TAN, performance changes depending on search algorithm choice is analyzed and evaluated in this study.

2. K2 and TAN

K2 algorithm is one of the special algorithms used for performance improvement in Bayesian network, and a technique that can effectively optimize each node. It proceeds by numbering node which is expressed with given node, and process each node in order. Regarding connecting of a line from previous node to present node, it adds by considering greedily and if present node cannot be optimized, it proceeds to a procedure that moves to next node [2]. Therefore, it is an algorithm that shows dependable result on the order of nodes that are set in the beginning. If it does not increase from parent node to present node, it is closed even if repeated. The performance needs to be verified, as K2 algorithm uses greedy technique. TAN (Tree Augmented Naïve Bayes) algorithm that extended from tree structure is a technique which expands Bayesian network, and a method that considers second parent node for each node, irrelevant to class nodes [4]. TAN model shows structure as following Figure 1.

3. Heart Disease and HRCT

Heart disease as one of the top causes for death in modern society is a dangerous disease which, stress that occurs in homes and works can be one of the cause. If heart disease occurs unexpectedly in everyday life, appropriate measure and quick treatment is important. If handling it is delayed, partial malfunction in body function or may face death if extreme. Heart diseases such as this can be prevented through continuously taking measures beforehand and minimizing its onset is most effective. Just like in reality, if heart disease risk can be predicted even in unclear situation due to numerous factors that affect, then it may positively affect managing of the disease and its prevention.

If applying purge, Bayesian network, and nerve network theory among data mining technique, even though the situation is unclear due to various affecting factor like in reality, probabilistic analysis is possible based on collected massive data [7, 8]. So in this study, it changes search algorithm in extended Bayesian network and evaluated the performance based on concluded result. Also, as risk of the patient’s heart disease risk in classified result can be verified, distinguishing high risk HRCT images is possible.

HRCT is a high resolution computer tomography, and one of the special techniques of CT. General CT has thickness of slice as about 5~10mm, and distance of 5~10mm which films twice consecutively. On the other hand, HRCT films by assigning slick thickness of about 1~3mm (usually 1.5~2mm) and increases clarity and resolution. It is mostly used to understand diagnosis and coverage of chronic lung disease or bronchiecstasia.

In order to show detailed lesion more clearly and truly in HRCT image, thinning the thickness of slice as much as possible is advantageous. However, there is a disadvantage of having to observe multiple filmed photos to observe slope running blood vessel or running of bronchial tubes, the thinner it gets. Since advantage of high clarity and resolution of HRCT image is greater than disadvantage, it may not be much of a problem once one becomes familiar with HRCT images

4. Experiment

4.1 Dataset Collection

WEKA v3.6.2 [6] developed in Waikato University is used as a tool for the experiment, and data used are composed of heart disease cases gathered from Cleveland Clinic Foundation of U.S., cleveland_data.arff. This experimental data contains information of actual patients, and is composed of 14 total attributes mixed with numeric and nominal type. There is age, trestbps, chol, thalach, oldpeak, ca, for numeric attribute, and sex, cp, fbs, restecg, exang, slope, thal, num exists for nominal attributes. The num data are probably value on classified result, that is written as one of the 5 types of {0, 1, 2, 3, 4} which means heart disease class. Total of 303 data are used for experimentation.

4.2 Data Preprocessing

The num attribute can be discretized to an integer between [0, 4], normal healthy state only when 0, and in other cases it means that one has a heart disease classified in each class. The unknown value and outlier that may exist in collected data can be corrected by taking median of similar cases or general means. But among experimental data attribute, num attribute identifies actual heart disease class which numerous factors need to be considered to decide, that makes it difficult to appropriately estimate. Therefore, less affecting in statistical result, general state of 0 is processed not to disturb result analysis. The affecting factor was minimized in experiment by revising attributes of remaining numeric types using the overall means of each attribute value.

4.3 Experimental Results

Based on data from cleveland_data.arff, selected num property, applied extended Bayesian network, and used K2 and TAN that is previously explained in search algorithm. In addition, fold value 10 of cross-validation was given and processed in data analysis. The Table 1 arranged measured correct and incorrect numbers on cases classified by total of 10 experiment result.

Table 1. 10-fold Validation of the Experiment

		1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*sum*	*avg*
Correct	K2	189	183	183	176	176	178	175	176	173	186	1795	179.5
	TAN	178	168	172	164	159	158	179	159	163	170	1670	167
Incorrect	K2	114	120	120	127	127	125	128	127	130	117	1235	123.5
	TAN	125	135	131	139	144	145	124	144	140	133	1360	136

Below Table 2 arranges and shows overall experiment result verified through 10 cross-validation in Bayesian network extended by K2 and TAN.

Table 2. K2 and TAN Applied to 10-folds CV Experiments

	K2	TAN
Correctly Classfied Instances	1795 (59.24 %)	1670 (55.12 %)
Incorrectly Classfied Instances	1235 (40.76 %)	1360 (44.88 %)
Kappa statistic	0.3564	0.2770
Mean absolute error	0.1754	0.2012
Root mean squared error	0.3389	0.3486
Relative absolute error	65.76 %	75.46 %
Root relative squared error	94.09 %	96.78 %
Total Number of Instances	3030	3030

5. Discussion of Experimental Results

Through experimental results of Table 2, similar result can be seen where MAE (Mean Absolute Error) of K2 is 0.1754, classification accuracy is about 59.24%, and MAE of TAN is 0.2012, classification accuracy is about 55.12%. In addition, following Figure 1 shows analysis on Kappa statistic by evaluation index of reproducibility on whether the experiment is similarly reproduced in case same experiment procedure in different view point is done.

If applying K2, average Kappa is 0.3564, and TAN shows average of 0.2770, which means that reproducibility of this experiment tend to be somewhat lacking. Figure 2 shows the square of error appearing in experiment and square of means RMSE (Root Mean Squared Error), which has similar meaning to standard deviation. In experiment result, average RMSE of 0.3389 is shown if K2 is applied, and average RMSE 0.3486 if TAN.

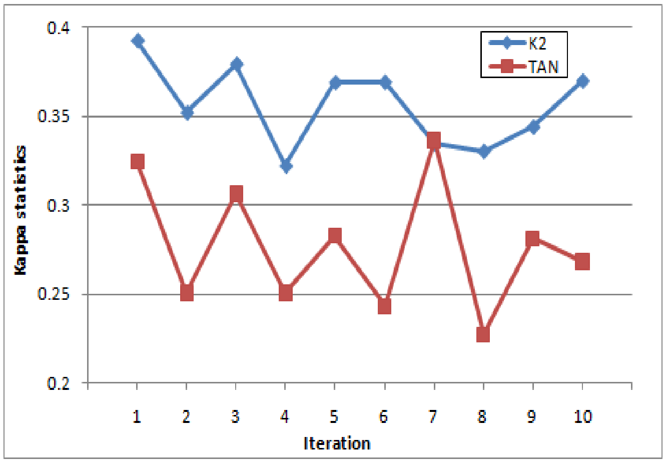

Figure 1. The Results of the Kappa Statistic

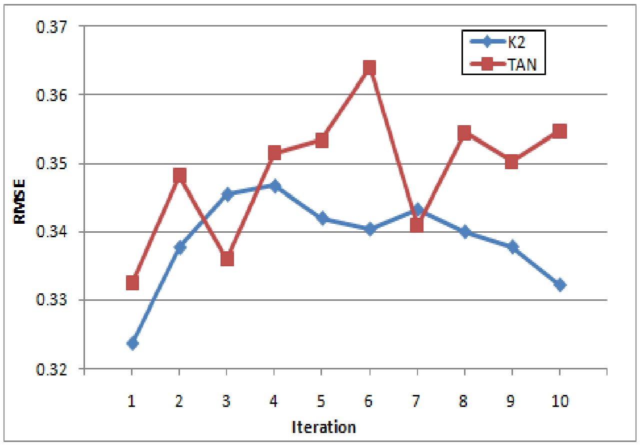

Figure 2. The Results of the RMSE

Bayesian network that is composed, based on probability approach on correlation of numerous factors may be used efficiently for reasoning heart disease risk of the patient. Similar to actual medical professional's decision process, intelligent reasoning by considering the human information and experience from previous case with the HRCT image can be done. Despite of existing experiment result that explains effectiveness of TAN compared to K2 in most cases, this experiment shows poor performance result of TAN overall. However clear conclusion on K2 is better algorithm than TAN cannot be done, yet indicates that more detailed experimentation is required. If RMSE and Kappa statistic is considered, the result which may mean coincident experiment due to low reproduce probability with potential error ratio can be interpreted as need for additional experimentation.

A case which TAN algorithm does not work better than K2 is verified through this experiment result. Multivariate data containing 14 attributes and experimental data with few cases process is presumed to be the main cause of this result. Compared to whole attribute numbers, low ratio of cases is studied, and as number of considerable numbers to verify the process is numerous, relatively low performance is shown.

Therefore, need to gather more cases of experiment data and apply of each algorithm in various view point is seen. In addition, method to improve performance of algorithm by excluding variables with low correlation in analyzing heart disease risk must be studied.

6. Conclusion

Recently, data mining field is becoming popular to extract information that can be used in diagnosis of actual patient, based on vast amount of data owned by the medical field. Additional medical images such as MRI, HRCT, and etc are used to complement lacking information.

Although various techniques such as association, clustering, Bayesian network which is algorithms modeling these information, no perfect algorithm is developed yet because of numerous factors to be considered exists in reality.

In this research heart disease related data collected from actual patients cleveland_data.arff has been classified by applying extended Bayesian network, and proceeded with the experiment while altering search algorithm of K2 and TAN respectively. Analysis on changing performance depending on search algorithm in extended Bayesian network is done, and high risk HRCT can be identified by using disease classification information verified from classification result.

For more accurate experimentation, performance will be evaluated by building environmental factors similar to reality in the future. Not only simple performance comparison will be taken but also complex application of numerous models through ensemble method such as bagging, boosting, and stacking will be studied. In addition, by marking the suspected area of actual disease location in HRCT image classified as high risk, the author is planning to make it possible to be used as helpful information for diagnosis.

References

[1] S. Moran, Y. Hey and K. Liu, “An Empirical Framework for Automatically Selecting the Best Bayesian Classifier”, Proceedings of the World Congress on Engineering 2009, vol. 1, **(2009)**.

[2] C. Ruiz, “Illustration of the K2 Algorithm for Learning Bayes Net Structures”, Department of Computer Science, WPI, **(2005)**.

[3] E. Lamma, F. Riguzzi and S. Storari, “Improving the K2 Algorithm Using Association Rule Parameters”, Modern Information Processing: From Theory to Applications B, **(2006)**.

[4] J. Davis, V. S. Costa, I. M. Ong, D. Page and I. Dutra, “Using Bayesian Classiers to Combine Rules”, Department of Biostatistics and Medical Informatics, University of Madison-Wisconsin, **(2004)**.

[5] J. Cerquides, “Tractable Bayesian Learning of Tree Augmented Naive Bayes Classifiers”, Ramon López de Màntaras, **(2003)**.

[6] I. C. Kim and Y. G. Jung, “Using Baysian Network to analyze Medical Data”, LNAI2734, Springer-Verlag, **(2003)**, pp. 317-327.

[7] Y. G. Jung, K. Y. Lee and M. J. Lim, “Discharge Decision for Post-Operative Patients”, Proceedings of ICHIT, **(2010)**, pp. 195-199.

[8] A. Mustafa, A. Akbar and A. Sultan, “Knowledge Discovery using Text Mining: A Programmable Implementation on Information Extraction and Categorization”, International Journal of Multimedia and Ubiquitous Engineering, vol. 4, no. 2, SERSC, **(2009)**, pp. 183-188.

[9] D. Bhattacharyya, S. Biswas and T. -h. Kim, “A Review on Natural Language Processing in Opinion Mining”, International Journal of Smart Home, vol. 4, no. 2, SERSC, **(2010)**, pp. 31-38.

[10] A. M. Khattak, A. M. Khan, S. Lee and Y. -K. Lee, “Analyzing Association Rule Mining and Clustering on Sales Day Data with XLMiner and Weka”, International Journal of Database Theory and Application, vol. 3, no. 1, **(2010)**, pp. 13-22.

Authors

Yong Gyu Jung received the B.S. in physics Education from Seoul National University in 1981. And then he got the M.S. and ph.D. degree of Computer Science from Yonsei and Kyonggi University in 1994 and 2003 respectively. Since 1999, he has joined as a Faculty of Eulji University in dept. of Medical IT marketing. His research interests are in the areas of medical information analysis and international standards including e-Business. He is a Member of ISO/TC154 standard organization. For more information, see http://dept.eu.ac.kr/mitm/overview/overview_teaching.asp.

Young-Jin Choi received Master of Business Administration, Hankuk University of Foreign Studies in 1988. And then he got Doctor of Business Administration, Sungkyunkwan University at 2004. Since 2006, he has joined as a Faculty of Eulji University in Dept. of Healthcare Management. His research interests are in the areas of IT Governance, Medical Information Systems. For more information, see http://dept.eu.ac.kr/S_Hospital/overview/overview_teaching.asp.