Evaluating the Quality of a Probabilistic Diagnostic System Using Different Inferencing Strategies

Yu-Chuan Li, M.D. and Peter J. Haug, M.D.
Department of Medical Informatics, University of Utah, Salt Lake City, Utah

In this paper we describe the evaluation of a probabilistic diagnostic system for patients with renal mass. Three inference models: Multi-membership Bayesian (MB), Minimal Diagnosis (MD) and Bayesian Network (BN), and 72 patients are used to illustrate three interrelated measures of system performance: accuracy, reliability and discriminating power. The inferencing strategies we tested demonstrated the kind of trade-offs in the performance measures that can be expected from imperfect systems. Ultimately, the purpose and expected use of a system should dictate the relative importance ascribed to different aspects of system performance.

INTRODUCTION

Medical diagnosis is one of the most ntellectually challenging processes in medical practice. Researchers have long attempted to eproduce this hypothetico-deductive process hrough the use of intelligent computer programs 1-4]. Thanks to vigorous research in this area, everal such programs have demonstrated hemselves as potentially useful tools in clinical onsultation, medical education, quality ssurance and clinical data capturing [2,3,5-11].

When constructing a medical diagnostic rogram, probabilistic models are often chosen n preference to rule-based or heuristically scored odels. Several advantages are associated with is approach. The use of probabilities allows e system to communicate succinctly the degree f certainty with which different diseases can be ssigned. It not only predicts the most likely agnosis, but also helps to clarify the relative stance between the most likely diagnosis and s competitors. In addition, probabilistic ediction of diagnoses provides information that n be used to quantitatively evaluate the ks/benefits (utility) of different work-up and erapeutic strategies [12]. However, before the obabilities produced by an expert system can be utilized in these ways, the quality of the probabilistic system has to be evaluated.

To evaluate a probabilistic system comprehensively, we believe that three interrelated parameters should be assessed, namely, *accuracy*, *reliability* and *discriminating power* [13,14]. Accuracy describes the ability of a system to assign the <u>highest</u> probability to the correct diagnosis. In this context, accuracy can be represented as the fraction of patients correctly diagnosed by the system or the non-error rate (NER).

By reliability we mean the trustworthiness of the probabilities suggested by the system. More specifically, how confidently can we translate these probabilities into the expected frequencies of the events. For example, when a reliable meteorologist says that there is a 80% chance of rain today, the implication is that rain will occur in eighty of a hundred days that have weather conditions similar to today.

Another important parameter apart from reliability and accuracy is the discriminating power, which represents the ability of a system to differentiate between likely and unlikely diseases. As an example of a non-discriminative system, consider a diagnosis suggestion list like "Acute myocardial infarction: 95%, Pulmonary embolus: 94%; Esophageal spasm: 92%". This formulation would not be helpful to the physician who is trying to differentiate these competing diagnoses.

We believe the three parameters discussed above are important when evaluating a probabilistic system because an inaccurate system that fails to predict diagnoses correctly is not only useless but misleading; an unreliable system hinders the generated probabilities from being used in decision-theoretic analysis and reduces its transferability, and a system that does not adequately separate the truly likely diagnosis from its less likely competitors can be confusing and, if believed, would carry the risk of increasing the number of tests as physicians sought to differentiate the diagnoses it marked as competitors [5,6].

In this study, we demonstrate the use of simple statistical tests to ascertain the quality of one diagnostic expert system running under three different inferencing strategies, namely, Multi-membership Bayesian (MB), Minimal Diagnosis (MD) and Bayesian Network (BN) [15-17]. The expert system used is a probabilistic renal mass diagnostic system (RMDS) that runs under the three models. To provide the data set necessary for these analyses, we recorded relevant findings from 72 renal mass patients. Accuracy, reliability and discriminating power were assessed for these models as the measures of system performance.

METHODS

Structure of The RMDS

The RMDS was developed using the ILIAD shell, which is a set of tools best known as the foundation of a large diagnostic system for internal medicine [3,7]. Diseases are constructed as frames in which the prior probabilities of the diseases and the conditional probabilities for findings are embedded. Several mechanisms including multi-level frames have been implemented in this system to handle conditionally dependent findings [18-21]. When only single-level frames are used in the knowledge base, the system behaves as a Multi-membership Bayesian program [18,19].

The construction of the RMDS uses principally the single-level structure. It consists of 18 probabilistic disease frames, each representing one category of renal mass (Table 1). The number of findings per frame ranges from 9 to 23 with an average of 15. Prevalence rates (prior probabilities) of the 18 renal mass diseases were calculated from a large patient database. The conditional probabilities for findings were estimated by two senior urologists. In a study to verify the validity of RMDS, the diagnostic accuracy for renal mass patients was compared between RMDS and six physicians. The result showed that RMDS performed better than the second-year residents and was comparable to chief residents trained in the urology department [22].

Table 1 The distribution of test cases into 18 renal mass diagnoses

Diagnosis	Number of cases
1) Renal parenchymal tumors	
Angiomyolipoma	4 (6%)
Hemangiopericytoma	0 (0%)
Juxtaglomerular cell tumor	0 (0%)
Lipoma	0 (0%)
Lymphoblastoma	4 (6%)
Metastatic tumor	2 (3%)
Oncocytoma	4 (6%)
Renal cell carcinoma	14 (19%)
Sarcoma	2 (3%)
Wilms' tumor	1 (1%)
2) Tumors of renal pelvis	
Benign papilloma	3 (4%)
Transitional cell carcinoma	17 (24%)
Squamous cell carcinoma	2 (3%)
Adenocarcinoma	2 (3%)
3) Renal cyst	
Simple cyst	4 (6%)
Cystadenocarcinoma	0 (0%)
4) Renal abscess	9 (13%)
5) Xanthogranulomatous pyelonephritis (XGP)	4 (6%)
Total	72 (100%)

Patients

Seventy-two consecutive cases of renal mass surgically diagnosed in the Chung Gung medical center between May 1989 and April 1992 were collected as our test cases. Findings from categories including basic demographic data, medical history, symptoms and signs, laboratory data and radiological diagnostic procedures were recorded and entered into the system. The final diagnoses of these cases were all confirmed by pathological examination and were used as the gold standard diagnoses in this study. The number of findings related to the renal mass diagnosis ranged from 14 to 30 (average 20) per case.

To explore the characteristics of the three models at different stages in the diagnostic workup, highly specific (and often expensive) examinations including renal angiogram, computed tomography scan and magnetic resonance imaging were removed from each case to produce a vignette. This vignette was labeled as "Phase 1" (an average of 1.8 findings were removed). Cases with a complete set of findings were then labeled as "Phase 2" vignettes.

The Three Inferencing Models

RMDS is based on a set of single-level frames constructed using the ILIAD shell. Two of the three inferencing models studied are standard strategies supported by the ILIAD shell. The first of these is a Multi-membership Bayesian model (MB). In the MB model, diseases are treated completely independently. Experience with such Multi-membership Bayesian diagnostic programs suggested that they frequently over-estimate the probabilities of diagnoses when trying to assess a set of competing diagnoses. Recent theoretic work suggests that this type of model does not truly reflect the joint distribution of the diseases and data in the system [16,17]. Instead, this model falsely assumes that the independence among diseases manifesting the same findings is maintained when those findings are known to be present. The result is that each disease receives all the information available to it from a shared finding. No finding can ever be thought of as "explained" by one member of a disease set. Thus, the finding "bone pain or tenderness" would still contribute a full compliment of evidence to "renal cell carcinoma" after "renal sarcoma" is proven.

Bayesian theorists describe the conditioned dependence of one disease on another (conditioned on a shared, instantiated finding) as "*d*-separation". In order to simulate the effects of *d*-separation in ILIAD, we have developed a model known as the "Minimal Diagnosis" (MD). This model selects and removes from the case a single high probability diagnosis which explains a large fraction of the patient findings. The remaining, "unexplained" findings produce a residual differential diagnosis designed to explain the remaining findings. This process was then repeated iteratively until all the important findings are attributed to particular disease hypotheses.

While the MD model provides one approach to *d*-separation, it represents an extremely aggressive way of assigning the information associated with clinical data. Bayesian Networks represent an alternative technique for propagating probabilities which is thought to handle *d*-separation accurately. A Bayesian Network, also called belief network or probabilistic causal network, is a graphical representation of probabilistic dependencies among variables [23]. More specifically, a Bayesian Network is a directed acyclic graph in which each node represents a random variable. The arrows in the graph often denote direct causal influences between variables, where the strength of the influence is specified by tables of conditional probabilities [16].

We were able to capitalize on structural similarities between ILIAD's native knowledge representation and Bayesian Networks to develop general tools for converting ILIAD knowledge base to coherent networks. We used these tools to convert the RMDS knowledge base from MB formulation into a BN formulation without any additional knowledge engineering effort. This BN formulation became the third model of RMDS in the performance comparison described below.

Measurement of Performance

As described in the introduction section, we have used accuracy, reliability and discriminating power as parameters for measuring the performance of these three inferencing strategies. The non-error rate (NER), which is the fraction of patients correctly diagnosed by the system, was used as the index of accuracy. McNemar's test for non-independent proportions was used to show the statistical significance of the NER's [24].

To measure reliability, we adopted a set of statistics described in detail by Habbema et al. regarding the measurement of performance in probabilistic diagnosis [13,14]. These statistics were arbitrarily denoted as Q1 through Q5. Q1 is the average of the probabilities (over all patients in the test population) that the program has assigned to each patient's correct diagnosis, while Q2 is defined as the <u>expected</u> value of this average. Q2 is derived from the probabilities assigned to all the diseases in all of the cases that have been processed by the system. The difference (Q1 minus Q2) between <u>observed</u> and <u>expected</u> mean diagnostic probabilities, called Q3, reflects the discrepancy between the computer's average estimate of the probability of the disease and the expected value. Apart from random fluctuations, Q3 averages zero for perfect reliable systems. If the sample size is not too small, the distribution of Q3 can be approximated by a normal distribution. When Q3 is divided by Q4, which is the standard deviation of the distribution of Q3, the result can be treated as a Z value from the standard normal distribution. This Z value is called Q5 in Habbema notation. 95% of sample values of Q5 from a perfectly reliable system should be within 1.96 standard deviations from zero. If the absolute value of Q5 is greater than 1.96, one must reject the null hypothesis that the program produces reliable probabilities. In this study, Q5 was used as the index of reliability.

The statistic we chose to represent discriminating power is called the *quadratic score* or Brier *score* [13,14,25,26]. The quadratic score combined two type of information; it uses the deviations of the probability assigned to the patient's real disease from 1.0 and the deviation of the probabilities assigned to the diseases that the patient does not have from 0 to produce a measure of discriminating ability. This value can be calculated for individual patients. The mean over a sample of test patients measures the system's overall discriminating ability. This score would be zero (perfectly discriminative) if the probability of correct diagnosis were always assigned 1 and those of the wrong diagnoses were assigned zero. In the other extreme, if the probability of the correct diagnosis were assigned zero and those of the wrong diagnoses were assigned 1 (the most unfavorable situation), this score would be equal to D, where D is the number of all possible diagnoses in the system. The larger the quadratic score, the less discriminative the system is. A repeated measure ANOVA was performed on the quadratic scores calculated from the probabilities assigned to each patient by the three models to determine the significance of differences in discriminating power.

RESULTS

Table 1 is a list of the diagnoses in RMDS and the distribution of the 72 test cases across these diagnoses. Patient age ranges from 6 to 80 with an average of 59. Thirty-four (47%) out of the 72 patients are female.

All of the 3 models (MB, MD and BN) were compared on the basis of an equivalent knowledge base and identical clinical information. Table 2 shows the values of NER, Q5 and quadratic score in the three models. In the analysis of accuracy, all models achieved NER accuracy greater than 61% in phase 1 and 69% in phase 2 (Figure 1a). No significant difference was found among the NER of MB, MD and BN in either Phase 1 or Phase 2. Phase 2 demonstrates a higher NER than Phase 1 because more information was used in Phase 2.

In the analysis of reliability, MB in phase 2 and MD in both phases resulted in absolute values of Q5 that were greater than 1.96 ($|Q5| > 1.96$), and thus should be treated as producing unreliable probabilities (Figure 1b). According to the values of Q5, only the BN model was able to generate reliable probabilities in both phases.

Using quadratic score as an index of discriminating power, MD and BN both showed better discriminating power (lower quadratic scores) than MB in both phases (Figure 1c). When we analyzed the quadratic scores using ANOVA, both MD and BN were found to have significantly smaller scores than MB in Phase 1 ($P< 0.02$) and Phase 2 ($P < 0.0001$), while no significant difference was found between MD and BN in either phase.

Table 2 The results of non-error rate (NER), Q5 and quadratic score for the three inference models (see also Figure 1)

Accuracy (NER)

	MB	MD	BN
Phase1	0.639	0.611	0.611
Phase2	0.764	0.694	0.722

Reliability (Q5)

	MB	MD	BN
Phase1	-1.70*	-3.20	0.91*
Phase2	-2.93	-6.56	-1.39*

Discriminating Power (quadratic score)

	MB	MD	BN
Phase1	0.872	0.644	0.681
Phase2	1.299	0.580	0.484

MB: Multi-membership Bayesian; MD: Minimal Diagnosis; BN: Bayesian Network
* $|Q5| < 1.96$

Bar charts of NER, Q5 and quadratic scores for Multi-membership Bayesian (MB), Minimal Diagnosis (MD) and Bayesian Network (BN) model across Phase 1 (Ph 1) and Phase 2 (Ph 2). (See Table 2 for the exact values)

Figure 1a Bar chart of the accuracy measure (NER (non-error rate); optimal: 1.0)

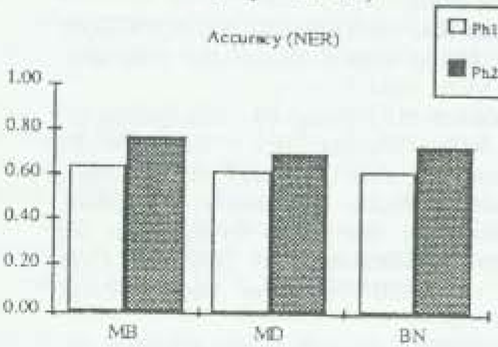


Figure 1b Bar chart of the reliability measure (optimal: -1.96 ~ +1.96)

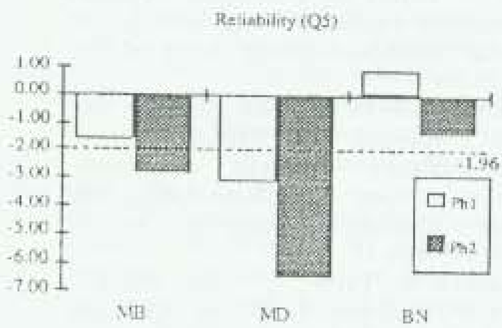
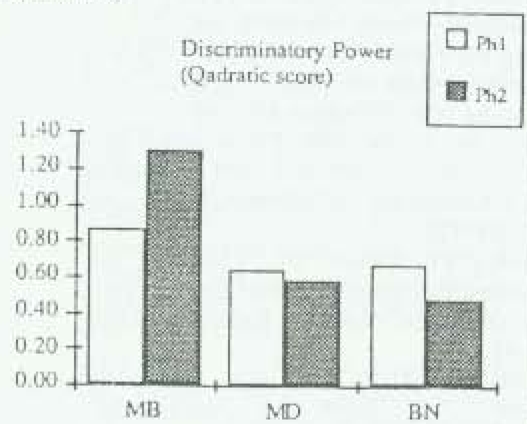


Figure 1c Bar chart of the discriminating power (optimal: 0)



DISCUSSION

Evaluation of clinical decision support systems is a complex issue [27,28]. This paper focuses on the analysis of system behavior using performance statistics. We believe that when evaluating a probabilistic diagnostic system, reliability and discriminating power are as important as the accuracy of the system.

In this study, these three parameters were assessed to evaluate a probabilistic renal mass diagnostic system under three models of inference: Multi-membership Bayesian, Minimal Diagnosis and a Bayesian Network formulation. All the models performed comparably in the test of accuracy. The MB model, our original implementation of RMDS, failed to pass the reliability test in Phase 2 and showed poor discriminating power in both phases. MD, by employing an aggressive *d*-separation algorithm, achieved better discriminating power than MB, yet sacrificed the ability to generate reliable probabilities in both phases. Although BN excelled in the tests of reliability and discriminating power over the other two models, the complexity of its algorithm made inferencing much slower than the other models. It took an average of 30 seconds to run a case in our 300 nodes Bayesian Network on a MacIntosh IIci computer. This compares to 2 seconds for the other two models on the same platform.

Based on these results, none of the three models has been shown to be perfect. The goals of a given implementation must, therefore, dictate the approach chosen. Among the models we evaluated, MB is not suitable for the applications where differentiating competing diagnoses is crucial because of its low discriminating ability. The probabilities generated by the MD model, due to their unreliability, should not be used as a source for decision-theoretic analysis where probabilities are treated as expected frequencies. But since both MB and MD exhibit short response time, either one could be used in applications where quick response is important, such as an on-line consulting system, or they could be used together as complementary parts of one system. The BN model, on the other hand, is most useful when all the qualities measured are required and where immediate response is not necessary. An example might be a quality control application.

Besides the overall better performance on reliability and discriminating power, Bayesian networks can also generate the expected probabilities of unknown findings. This ability has not been studied in this experiment but could potentially be applied in clinical information acquisition and in clinical predictions. However, in spite of the many potential benefits of BN model, the exact inferencing algorithms for Bayesian networks, such as the one we are using, are computational intensive [17]. The number of calculations needed in these algorithms increases exponentially with the size and complexity of the network and thus very large Bayesian networks are deemed computational intractable [29]. Fortunately, recent research regarding inexact inferencing algorithms may lead to a solution for this problem [30-33].

The RMDS we evaluated in this study is a very narrowly defined diagnostic system that only deals with patients having certain forms of renal mass. The relatively small sample size and the rarity of some of the diseases in this system rendered analyses on individual disease categories inaccessible. This procedure could be of value in identifying disease-specific problems in a system.

Given the architecture of the ILIAD shell, where frames can be built separately and combined together to form larger systems, it is tempting to try assessing the performance statistics in a more generic system with a larger sample of patients. However, it would be challenging to obtain a large enough representative sample of cases for the evaluation of such a large system. In addition, the current algorithm used for the BN model may not be suitable for much larger systems because of increasing computation time.

A system that can generate probabilities of diagnoses is appealing because of the versatility of probability itself. But, if the probabilities generated from such a system lack reliability and discriminating power, they mean no more than a list of ranking scores. The three parameters of probabilistic systems described in this paper can be treated as indices for the usability of such probabilities. However, as in the case of the models in this study, perfect performance in all of these parameters may not be easily attainable. Ultimately, the purpose and expected use of a system should dictate the relative importance ascribed to different aspects of system performance.

* This publication was supported in part by grant number 5 R01 LM05323 from the National Library of Medicine.

References

[1]. Shortliffe EH. Computer programs to support clinical decision making. JAMA. 1987;258:61-6.

[2]. Miller RA. INTERNIST-1/CADUCEUS: Problem facing expert consultant programs. Meth Inf Med. 1984;23:9-14.

[[3]. Warner HR, Haug PJ, Bouhaddou O, Lincoln MJ, Warner HRJ, Sorenson D, Williamson JW, Fan C. Iliad as An Expert Consultant to Teach Differential Diagnosis. Proceedings of the 12th Symposium on Computer Applications in Medical Care (SCAMC), IEEE Computer Society Press 1988:371-376.

4]. Barnett GO, Cimino JJ, HuppJA, Hoffer EP. DXplain: An evolving diagnostic decision-support system. JAMA. 1987;258:67-74.

[5]. Bankowiz RA, McNeil MA, Challinor SM, Miller RA. Effect of a computer-Assisted general medicine diagnostic consultation service on house staff diagnostic strategy. Meth Inf Med 1989; 28:352-6.

[6]. Bankowiz RA, McNeil MA, Challinor SM, Parker RC, Kapoor WN, Miller RA. A computer-assisted medical diagnostic consultation service - Implementation and prospective evaluation of a prototype. Ann Int Med 1989; 110:824-32.

[7]. Cundick R, Turner CW, Lincoln MJ, Buchanan JP, Anderson C, Warner HRJ, and Bouhaddou O. Iliad as a patient case simulator to teach medical problem solving. Proceedings of the 13th Symposium on Computer Applications in Medical Care (SCAMC), IEEE Computer Society Press, 1989:902-906.

[8]. Haug PJ, Frederick PR, Tocino I. Quality control in a medical information system. Med Decis Making 1991;11(suppl):S57-S60.

[9]. Lau LM, Warner HR. Performance of a Diagnostic System (Iliad) as a Tool for Quality Assurance. Computers and Biomedical Research 1992; 25:314-323.

[10]. Haug PJ, Ranum DL, Frederick PR. Computerized extension of coded findings from free-text radiologic reports. Radiology 1990; 174:543-48.

[11]. Haug PJ, Clayton PD, Tocino I, Morrison JW, Elliot CG, Collins DV, Harada SK, Frederick PR. Chest radiography: A tool for the audit of report quality. Radiology 1991; 180:271-76.

[12]. Sox HC, Blatt MA, Higgins MC, Marton KI. Medical Decision Making. Butterworth-Heinemann, Boston 1988.
[13]. Habbema JDF, Hilden J, Bjerregaard B. The measurement of performance in probabilistic diagnosis (parts 1, 2, and 3). Meth Inf Med 1978; 17:217-46.
[14]. Habbema JDF, Hilden J, Bjerregaard B. The measurement of performance in probabilistic diagnosis (parts 4 and 5). Meth Inf Med 1981; 20:80-100
[15]. Ben-Bassat M, Carlson RW, Puri VK, Davenport MD, Schriver JA, Latif M, Smith R, Portigal LD, Lipnick EH, Weil MH. Pattern-Based Interactive Diagnosis of Multiple Disorders: The Medas System. IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol.PAMI-2, No.2, March 1980.
[16]. Pearl J. Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference. Morgan Kaufman, San Mateo, CA, 1988.
[17]. Neopolitan E. Probabilistic Reasoning in Expert Systems. Wiley, New York, NY 1990
[18]. Yu H, Haug PJ, Lincoln MJ, Turner C, Warner HR. Clustered knowledge representation: Increasing the reliability of computerized expert systems. Proceedings of the 12th Symposium on Computer Applications in Medical Care (SCAMC), IEEE Computer Society Press 1988:126-130.
[19]. Turner CW, Lincoln MJ, Haug PJ, Warner HR, Williamson JW, Whitman N. Clustered disease findings: aspects of expert systems. International Symposium of Medical Informatics and Education. Salamon R, Protti D, Moehr J. eds. University of Victoria, B.C., Canada, 1989:259-63
[20]. Sorenson DK, Cundick RM, Fan C, Warner HR. Passing Partial Information among Bayesian and Boolean Frames. Proceedings of the 13th Symposium on Computer Applications in Medical Care (SCAMC), Washington DC: IEEE Computer Society Press 1989:50-54.
[21]. Lincoln MJ, Haug PJ, Hong Y, Turner C, Warner HR. Expert Biases Prevent Accurate Estimation of Population Statistics for Clustered Disease Frames. International Symposium of Medical Informatics and Education. Salamon R, Protti D, Moehr J. eds. University of Victoria, B.C., Canada, 1989:237-241.
[22]. Chung PL, Li Y, Wu CJ, Huang MH. Using Iliad system shell to create an expert system for differential diagnoses of renal masses. (unpublished)
[23]. Howard RA, Matheson JE. Inflence diagrams. In: Readings on the Principles and Applications of Decision Analysis. Howard RA, Matheson JE. eds. Menlo Park, CA: Strategic Decisions Group, 1981:721-62.
[24]. McNemar Q. Psychological statistics, 4th edition. Wiley, New York. 1969:54-63.
[25]. Brier GW, Allen RA. Verification of Weather Forecasts. In: Compendium of Meteorology. Malone TF. eds. Boston: Amer Meteorol Soc, 1951:844-848.
[26]. Murphy AH. Evaluation of probabilistic forecasts: some procedures and practices. In: Weather Forecasting and weather forecasts. Murphy AH, Williamson DL. eds. Boulder Colorado: National Center for Atomspheric Research, 1977:807-30.
[27]. Hilden J, Habbema JDF. Evaluation of clinical decision aids-more to think about. Med Inform 1990;15:275-284.
[28]. Miller PL. Issues in the evaluation of artificial intelligence system in medicine. Proceedings of the Ninth Annual Symposium on Computer Applications in Medical Care. New York: IEEE Computer Society Press; 1985;281-6.
[29]. Cooper GF. The Computational Complexity of Probabilistic inference using Bayesian Belief Networks. Artificial Intelligence 1990; 42:393-405.
[30]. Henrion M. Toward Efficient Probabilistic Diagnosis in Multiply Connected Belief Networks. In: Influence Diagrams, Belief Nets, and Decision Analysis. Oliver RM, Smith JQ. eds. Wiley, Chichester, 1990:385-407.
[31]. Fung R, Chang KC. Weighting and Integrating Evidence for Stochastic Simulation in Bayesian Networks. In: Machine Intelligence and Pattern Recognition: Uncertainty in Artificial Intelligence 5, Vol 10. Herion M, Shachter R, Kanal LN, Lemmer JF. eds. Elsevier, Amsterdam, 1990:209-20.
[32]. Shachter RD, Peot M. Simulation Approaches to General Probabilistic Inference on Belief Networks. In: Machine Intelligence and Pattern Recognition: Uncertainty in Artificial Intelligence 5, Vol 10. Herion M, Shachter R, Kanal LN, Lemmer JF. eds. Elsevier, Amsterdam, 1990:221-31.
[33]. Henrion M. Propagating Uncertainty in Bayesian Networks by Probabilistic Logic Sampling. In: Uncertainty in Artificial Intelligence. Kanal LN, Lemmer JF. eds. Elsevier, Amsterdam, 1986:47-67.