A New Measure for the Accuracy of a Bayesian Network

Alexandros Pappas, Duncan Gillies

Department of Computing, Imperial College of Science, Technology and Medicine, 180 Queen's Gate, London SW7 2BZ, United Kingdom
{ap297, dfg}@doc.ic.ac.uk

Abstract. A Bayesian Network is a construct that is used to model a given joint probability distribution. In order to assess the quality of an inference, or to choose between competing networks modelling the same data, we need methods to estimate the accuracy of a Bayesian network. Although the accuracy of a Bayesian network can be easily defined in theory, it is rarely possible to compute it in practice for real-world applications due to the size of the space representing the variables. Instead, alternative characteristics of a Bayesian network, which relate to and reflect the accuracy, are used. A popular formalism that adopts such methods is the Minimum Description Length (MDL). It models the accuracy of a Bayesian network as the probability of the Bayesian network given the data set that it models. However in the context of Bayesian Networks, the MDL formalism is flawed, exhibiting several shortcomings. In its place, we propose a new framework for Bayesian Networks. We specify a measure, which models the accuracy of a Bayesian network as the accuracy of the conditional independencies implied by its structure. Experiments have been conducted, using real-world data sets, to compare MDL and the new measure. The experimental results demonstrate that the new measure is much better correlated to the real accuracy than the MDL measure. These results support the theoretical claims, and confirm the significance of the proposed framework.

1 Introduction

A Bayesian network is a construct that is used to model a given joint probability distribution. In effect, a Bayesian network represents the relationships between a set of variables [8, 10]. A Bayesian network does not necessarily represent causation between a set of variables. For example, the parents of a node should not necessarily be considered causes of the node, although they can be interpreted as such in certain cases; instead, they should be viewed as shields against other influences.

A principal feature of a Bayesian network is the conditional independencies between the variables implied by the structure of the network. The absence of an arc (direct connection) between two nodes of a Bayesian network implies certain conditional independencies regarding these nodes. However, the structure of a Bayesian network does not imply dependencies between the variables of the joint probability distribution that it represents. The existence of an arc (direct connection) between two nodes of a Bayesian network does not imply dependence regarding these

nodes. All independencies implied by the structure of a Bayesian network are considered conditional. An unconditional independence can be viewed as a conditional independence given the empty set.

In this paper, we make three assumptions about the Bayesian network, which limit the scope of the research. Firstly, the Bayesian network models a data set. We are not concerned with networks that are estimated by subjective methods. Secondly, we assume that the data set is complete, and lastly that the variables of the data set are finite.

2 The Accuracy of a Bayesian Network

We will start with an intuitively appealing definition of the accuracy of a Bayesian network. A Bayesian network (*BN*) is accurate with respect to a data set (*D*), if and only if, the joint probability distribution represented by the Bayesian network (P_{BN}) matches the joint probability distribution described by the data set (P_D).

Both the joint probability distribution represented by the Bayesian network and the joint probability distribution described by the data set can be represented as *n*-dimensional matrices of $states_1 * \ldots * states_n$ elements, assuming the data set has *n* variables, and each variable v_i has $states_i$ states. Since a matrix of *n* elements can be represented geometrically as a point in the $\Re^n$ space, then both the joint probability distribution represented by the Bayesian network and the joint probability distribution described by the data set can be represented geometrically as points in the $\Re^{states_1 * \ldots * states_n}$ space. In view of the geometrical representation, the degree to which the joint probability distribution represented by the Bayesian network matches the joint probability distribution described by the data set is reflected by how close the corresponding points are.

The degree of accuracy of a Bayesian network is inversely related to the distance between the point corresponding to the joint probability distribution represented by the Bayesian network and the point corresponding to the joint probability distribution described by the data set.

Correspondingly, the degree of inaccuracy of a Bayesian network, with respect to a data set, is defined as the geometrical distance between the point corresponding to the joint probability distribution represented by the Bayesian network and the point corresponding to the joint probability distribution described by the data set.

$$inaccuracy(BN) = inaccuracy(BN, D) = distance(P_{BN}, P_D) \quad (1)$$

The Euclidean distance is used as a distance measure; alternative distance measures could be used, such as the L1 and L2 metrics.

$$inaccuracy(BN) = \sqrt{\sum_{v_1=1}^{states_1} \ldots \sum_{v_n=1}^{states_n} (P_{BN}(v_1, \ldots, v_n) - P_D(v_1, \ldots, v_n))^2} \quad (2)$$

The joint probability distribution described by the data set is determined directly from the data set, while the joint probability distribution represented by the Bayesian

network is determined indirectly from the data set. The range of values for the degree of inaccuracy of a Bayesian network is $[0,\sqrt{2}]$.

Unfortunately, although the accuracy of a Bayesian network is well defined in theory, and an appropriate measure of inaccuracy is specified, it is rarely possible to determine the degree of inaccuracy of a Bayesian network in practice for real-world applications. This is due to the fact that in most cases it is computationally unfeasible to determine and use the joint probability distribution described by the data set, because of both processing and storage limitations. For example, for the Hepatitis C data set, which was used for our experimentation, the matrix of the joint probability distribution described by the data set is 9 dimensional containing 14,696,640 elements. So, it is not only computationally unfeasible to determine the matrix, but also virtually impossible to use the matrix in practice. This is principally the reason why a Bayesian network is used to model the joint probability distribution described by the data set, instead of the joint probability distribution itself.

3 Minimum Description Length (MDL)

We noted in the previous section that it is rarely possible to determine directly the degree of accuracy of a Bayesian network. For real-world applications, alternative characteristics of a Bayesian network, which relate to and reflect the accuracy, are used to model and examine the accuracy of a Bayesian network. A formalism commonly used for this purpose is the Minimum Description Length (MDL) [11] (see also [4]). The MDL formalism evaluates a model of a data set based on the length of the description of the data set. This is the sum of the length of the description of the model and the length of the description of the data set given the model. The length of the description of the model reflects the model size and complexity, while the length of the description of the data set given the model is interpreted as the model accuracy.

In the context of the Communication domain, the MDL formalism is employed in data compression, in order to identify the model that provides the shortest description of the data set. In this case, the length of the description of the data set is the number of bits required to encode the data set. This is the sum of the number of bits required to describe the model and the number of bits required to encode the data set given the model.

In the context of Bayesian Networks, the length of the description of the data set given the Bayesian network is the negation of the log likelihood of the Bayesian network given the data set, which is interpreted as the degree of inaccuracy of the Bayesian network. In effect, the MDL formalism models the accuracy of a Bayesian network with respect to a data set as the likelihood of the Bayesian network given the data set.

$$accuracy(BN) = accuracy(BN, D) = \log_2(P(BN \mid D)) \quad (\mathbf{3})$$

The likelihood of a Bayesian network given the data set is equivalent to the likelihood of the data set given the Bayesian network.

$$\max_{BN\in B}\{accuracy(BN,D)\} \overset{MDL}{=} \max_{BN\in B}\{P(BN \mid D)\} =$$
$$\max_{BN\in B}\{\frac{P(D \mid BN)P(BN)}{P(D)}\} = \max_{BN\in B}\{P(D \mid BN)P(BN)\} = \quad (4)$$
$$\max_{BN\in B}\{P(D \mid BN)\}$$
$$\text{since } \forall BN \in B, P(D) \text{ is constant, and } P(BN) \text{ is uniform}$$

The range of values for the degree of accuracy of a Bayesian network, with respect to any data set, is (-∞,0].

Since the MDL formalism evaluates the likelihood of a Bayesian network given a particular data set, the specific range of values for the degree of accuracy of a Bayesian network, with respect to a given data set, depends on the nature of the data set. As a result, it is not possible to determine the degree of accuracy of a Bayesian network unless the nature of the data set is examined. The degree of accuracy of a Bayesian network, with respect to a data set, is affected directly by the size of the data set [3]. Consequently, the MDL formalism can only be used to compare "relative" degrees of accuracy for a set of Bayesian networks and for a particular data set. Any results acquired and any conclusions drawn are valid for the Bayesian networks, only in view of the particular data set.

Although the MDL formalism is being used in the field of Bayesian Networks, it was initially developed for the Communication domain where the focus is the transmission of a message. In the field of Bayesian Networks, the focus is the construction of a network that models the joint probability distribution in a given data set. Evidently, the semantics of these two fields are different, and thus the MDL formalism, which has been developed for the communication domain, is being taken out of context when used in the field of Bayesian Networks.

The MDL formalism examines the accuracy of a Bayesian network with respect to the data entries of the data set, and not with respect to the joint probability distribution represented by the data set.

Other formalisms that have been developed to evaluate the characteristics of a Bayesian network are the Akaike Information Criterion (AIC) [1] and the Bayesian Information Criteria (BIC) [14]. However, both of these formalisms are identical to the MDL formalism with regards to the evaluation of the accuracy of a Bayesian network. Bayesian Network learning algorithms, such as the Maximum Weight Spanning Tree (MWST) algorithm [2], provide some sort of an evaluation scheme for a Bayesian network. However, each algorithm is based on its own heuristic methods, which do not specify clearly what the characteristics of a Bayesian network are, and how these are evaluated.

4 A New Definition of Accuracy in Bayesian Networks

We have developed a formal framework for describing the accuracy of Bayesian Networks. Full details will be found in [9], but for now we state just the central theorems of the framework.

Given:

- A data set (D) representing a joint probability distribution (P) over variables (V)
  $P_D = P(V)$
- A Bayesian network (BN) of nodes (N), arcs (A), distribution (P')
  $BN = (N, A, P')$
  $P_{BN} = P'(N)$

Provided:

- The nodes of the Bayesian network are the variables of the joint probability distribution represented by the data set
  $N \equiv V$
  $\forall i, n_i \in N, v_i \in V : n_i \equiv v_i$
- The probabilities of the Bayesian network reflect the corresponding probabilities of the joint probability distribution represented by the data set
  $\forall i, n_i \in N : P'(n_i \mid parents(n_i)) = P(n_i \mid parents(n_i))$

Then:

- The Bayesian network is accurate with respect to the data set, if and only if, the conditional independencies implied by the structure of the Bayesian network are conditional independencies of the joint probability distribution represented by the data set
  $P_{BN} = P_D \Leftrightarrow \forall i, v_i \in V, \forall W, W \subseteq a(v_i) : P(v_i \mid W \cup parents(v_i)) = P(v_i \mid parents(v_i))$

We refer to the set of conditional independencies implied by the structure of a Bayesian network as the Network Conditional Independencies (NCI) set. Our framework introduces the NCI Soundness theorem and the NCI Incompleteness theorem, which are remarkably significant for the field of Bayesian Networks.

The NCI Soundness theorem indicates that given a data set there exists a Bayesian network whose NCI is sound, that is the network conditional independencies also belong to the data conditional independencies. It guarantees that given any data set it is possible to construct an accurate Bayesian network.

The NCI Incompleteness Theorem indicates that there exists data sets (at least one) for which there exist no Bayesian networks whose NCI is complete. Thus given any data set it might be impossible to construct a Bayesian network whose structure implies all the conditional independencies of the joint probability distribution represented by the data set.

Computationally it is feasible to measure the inaccuracy of each conditional independence implied by the network, using a measure such as Mutual Information. The conditional dependency measure (DM) for variables A and B given variable C is the conditional Mutual Information (MI) for variables A and B given variable C.

$$DM(A, B \mid C) = MI(A, B \mid C) = \sum_A \sum_B \sum_C P(A, B, C) \log_2 \frac{P(A, B \mid C)}{P(A \mid C) P(B \mid C)} \quad (5)$$

By setting $C = \varnothing$ we obtain an unconditional dependency measure. Using the framework theorems and the definition of Mutual Information, it is possible to derive the following definition for the inaccuracy of a Bayesian Network.

$$
\begin{aligned}
&inaccuracy(BN) = inaccuracy(NCI) \\
&inaccuracy(NCI) = \begin{cases} \sum_i inaccuracy(NCI_i) & NCI \neq \varnothing, NCI_i \in NCI \\ 0 & NCI = \varnothing \end{cases} \\
&inaccuracy(NCI_i) = DM(NCI_i) \\
&DM(NCI_i) = MI(NCI_i)
\end{aligned} \tag{6}
$$

We call this measure the Network Conditional Independencies Mutual Information (NCIMI) measure.

The framework facilitates the comparison of individual Bayesian networks with regards to their characteristics, and provides the theoretical means to justify when and why a particular Bayesian network is more accurate than another Bayesian network. The framework can clarify the procedure of the addition and deletion of arcs and nodes from the structure of a Bayesian network [13], and provide the means to illustrate and explain the effects of such actions on the characteristics of a Bayesian network. The framework can supply a theoretical rationale to the process of the introduction of a hidden node within the structure of a Bayesian network [7] and the effects of such an action. The framework can be employed to develop Bayesian Network construction (learning) algorithms [15] as tree search algorithms that examine the accuracy of a Bayesian network and use the degree of inaccuracy as the evaluation function of the tree nodes.

5 Experimental Results

Several experiments were conducted using real-world problems. One data set that was used represents information concerning Hepatitis C patients, while a second data set provides information about morphological development of neurons [6]. The Hepatitis C data set has 1672 data entries, each being an individual Hepatitis C patient characterised by 9 distinct variables. The neurons data set has 44 data entries, each about an individual neuron, characterised by 6 distinct variables. At any time, only three variables of the given data set were used to carry out the experiments and collect experimental results. This limited approach was adopted, so that it was computationally feasible to calculate the joint probability distribution represented by the data set. The experiments were conducted for a collection of different data sets, created by randomly selecting the set of variables. For a given data set, all possible tree structured Bayesian networks that can model the given data were used to carry out the experiments and collect experimental results.

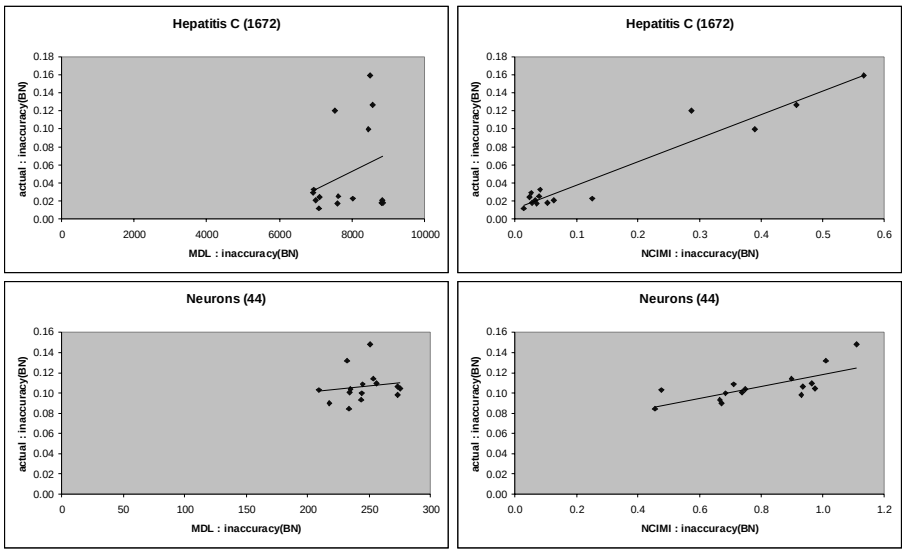
For a given data set and a given Bayesian network, the experiments determine the accuracy of the Bayesian network with respect to the data set. Three quantities were computed. First, the actual degree of inaccuracy of the Bayesian network, computed as the Euclidean distance between the data distribution and the joint probability distribution of the network. Secondly, the degree of inaccuracy of the Bayesian network according to the Minimum Description Length formalism, and thirdly the degree of inaccuracy of the Bayesian network according to our new measure.

The experimental results are presented below in a tabular and graphical form. The graphs are scatterplots of the actual degree of inaccuracy of the Bayesian network, against either the degree of inaccuracy measured by the MDL, or our NCIMI measure. Trendlines are drawn and their corresponding R^2 value computed. The closer the R^2 value of the trendline is to 1, the more reliable the trendline is, and the better it fits the points of the graph. The best trendline for the points of a graph is associated with the correlation of the variables represented in the graph. A low correlation coefficient indicates that the variables are unrelated.

According to the MDL formalism, the degree of inaccuracy of a Bayesian network, with respect to a data set, is affected directly by the size of the data set [3]. For the data sets of the Hepatitis C domain, the average degree of inaccuracy of the Bayesian networks examined is 7858.3; for the data sets of the Neurons domain, the average degree of inaccuracy of the Bayesian networks examined is 245.26. Thus, the average degree of inaccuracy of the Bayesian networks for the data sets of the Hepatitis C domain is 32.04 times greater than the average degree of inaccuracy of the Bayesian networks for the data sets of the Neurons domain. This result is not only of the same magnitude but also very close to the size ratio of the Hepatitis C data set and the Neurons data set, which is 38. The experimental results point to a linear correlation between the degree of inaccuracy of a Bayesian network, with respect to a data set, and the size of the data set, which agrees with what is predicted in theory.

The NCIMI measure does not exhibit the shortcomings of the MDL formalism. It provides a proper error norm so that the inaccuracy of the networks for the different data sets can be compared. It will be seen both visually and by the statistical measures, that it reflects the true accuracy of the network much better than MDL.

	Correlation (Hepatitis C)	**Correlation (Neurons)**	**R^2 value (Hepatitis C)**	**R^2 value (Neurons)**
MDL	0.31	0.14	0.09	0.02
NCIMI	0.97	0.73	0.94	0.53



6 Conclusion

We have proposed a framework for Bayesian Networks, in order to examine the accuracy of a Bayesian network. The framework has subtle similarities with research by Pearl [10], Neapolitan [8], and Chow & Liu [2]. Further investigation is required to identify these similarities, and to amalgamate these seemingly different methodologies into a unified framework. Using the framework, we have proposed a new accuracy measure for Bayesian networks called the NCIMI measure, which is based on assessing conditional independencies implied by the structure of the Bayesian network. The framework is formally established, with several definitions and theorems, and well-defined semantics. Applying our measure to real world problems we have demonstrated that it performs considerably better than the popular MDL measure, which was defined for a different application within a different formal system and with a different set of assumptions.

The data sets employed for the experiments are data sets of just three variables. Supplementary experiments are in hand, employing larger data sets with more than three variables. However, the data sets have to remain relatively small so that the experiments are computationally feasible. The refinement and the extension of the framework will result in an even greater understanding of Bayesian Networks, while the supplementary experiments will offer further insight, and additional evidence in support of the proposed framework.

References

1. Akaike H. “A new look at the statistical identification model”, IEEE Transactions on Automatic Control, 19:716-723, 1974
2. Chow C.K., Liu C.N. “Approximating discrete probability distributions with dependence trees” IEEE Transactions on Information Theory, 14:462-467, 1968
3. Friedman N., Yakhini Z. “On the sample complexity of learning Bayesian networks” Proceedings of the 12th Annual Conference on Uncertainty in Artificial Intelligence, 1996
4. Friedman N., Geiger D., Goldszmidt M. “Bayesian network classifiers” Machine Learning, 1997
5. Jensen F.V. “An introduction to Bayesian networks” UCL Press, 1996
6. Kim J., Gillies D.F. “Automatic Morphometric Analysis of Neural Cells” Machine Graphics and Vision 7(4), 1998
7. Kwoh C.K., Gillies D.F. “Using Hidden Nodes in Bayesian Networks” Artificial Intelligence 88:1-38, 1996
8. Neapolitan R.E. “Probabilistic reasoning in expert systems: theory and algorithms” Wiley-Interscience, 1990
9. Pappas A., Gillies D. “The Accuracy of a Bayesian Network” Technical Report, Imperial College, 2001 (available from the authors)
10. Pearl J. “Probabilistic reasoning in intelligent systems: networks of plausible inference” Morgan Kaufmann, 1988 (4th printing, 1997)
11. Rissanen J. “Modelling by shortest data description” Automatica, 14:465-471, 1978
12. Russell S., Norvig P. “Artificial Intelligence: a modern approach” Prentice Hall International, 1995

13. Sucar L.E., Gillies D.F., Gillies D.A. “Uncertainty Management in Expert Systems” Artificial Intelligence 61:187-208, 1993
14. Schwarz G. “Estimate the dimension of a model” The Annals of Statistics, 6(2):461-464, 1978
15. Tahseen T. “A new approach to learning Bayesian network classifiers” Ph.D. Thesis, Imperial College, 1998