J.Natn.Sci.Foundation Sri Lanka 2013 41 (3): 203-208
DOI: http://dx.doi.org/10.4038/jnsfsr.v41i3.6054

RESEARCH ARTICLE

A novel mutual dependence measure in structure learning

Muhammad Naeem* **and Sohail Asghar**
[1] *Department of Computer Science, Mohammad Ali Jinnah University, Islamabad, Pakistan.*
[2] *University Institute of Information Technology, PMAS-Arid Agriculture University, Rawalpindi, Pakistan.*

Revised: 29 January 2013; Accepted: 19 April 2013

Abstract: Mutual dependence between features plays an important role in the formulation of classifiers, clustering and other machine intelligent techniques. In this study a novel measure of mutual information known as integration to segregation (I2S), explaining the relationship between the two features is proposed. Some important characteristics of the proposed measure was investigated and its performance in terms of class imbalance measures was compared. It was shown that I2S possesses the characteristics, which are useful in controlling overfitting problems. In structure learning techniques such as Bayesian belief networks, conventional measures of dependency relationship cope with the overfitting problem by restricting the number of parents for a node; however it is still not impressive because complete overfitting is not eliminated. In contrast, I2S is capable of significantly maximizing the discriminant function with a better control of overfitting in the formulation of structure learning.

Keywords: Asymmetric, information theory, mutual dependence, structure learning.

INTRODUCTION

Various computational techniques have produced large amounts of data dealing with multifarious complexities and noticeable heterogeneity, yielding uncertainties and risks. Machine learning and data mining techniques have enabled the researchers to extract useful patterns out of a large dataset. Classification is a notable and impressive technique in machine learning and data mining. A classifier can be defined as a function $f : D_x \rightarrow D_c$ where class instance is defined as $c \in D_c = \{c_1, c_2, \ldots c_m\}$ to the objects described by a set of attributes $X = \{x_1, x_2, \ldots x_n\} \in D_x$. The dataset of attributes X contain N labelled instances of $< X, C >$ with the objective of correctly predicting the class label of a new data instance in the learning phase of a classifier. Among many of the classification systems introduced, a Bayesian belief network (BBN) is considered a robust technique by virtue of its ability to decompose complex probabilistic models into brief and tractable elements (Jensen & Neilson, 2007). The data mining community has extensively used it in knowledge discovery tools due to its solid statistical foundation and the capability for inference (Cooper & Herskovits, 1992; Chen *et al.*, 2008; Etminani *et al.*, 2010; Carvalho *et al.*, 2011). The BBN is a strong probabilistic model for knowledge representation.

A BBN is drawn by a directed acyclic graph (DAG) representing a set of conditional probability distributions for each stochastic node of the DAG; whereas, each arc between the two nodes represents the direction of inference or induction. A node (child), which is directly pointed to by another node (parent) receives inference from its parent node(s), while the parent node obtains induction from the child node in terms of probabilistic distribution. These concepts of inference and induction are helpful in formulating BBN classifiers.

The mutual dependence and correlation between two attributes of a dataset is a key problem in the sphere of structure learning. Numerous pairwise measures have been introduced explaining a particular or general relationship (Gibbons & Subhabrata, 2003; Wasserman, 2007; Corder & Foreman, 2009; Bagdonavicius *et al.*, 2011). However, it has been described that correlation and dependence are intrinsically different phenomena. Although wide application of correlation in various domains of interest has been reported, a careful examination of correlation measures highlights two

* Corresponding author (naeems.naeem@gmail.com)

problems in structure learning. The first issue is related to its incapability of describing the nonlinear structure between the random variables. It has been pointed out that two uncorrelated variables do not suggest their independence to each other (Grimmett & Stirzaker, 2001). The second problem is the inability of providing circumscribed knowledge about the underlying true dependence nature (Grimmett & Stirzaker, 2001). Thus arises a dictum that "*correlation is unable to imply causation*" emphasizing that correlation is not well suited in classification problems for the sake of establishing causal relationships between variables (Aldrich, 1995). Jensen & Neilson (2007) elaborated two important characteristics for scoring functions used in the belief network; (a) the ability of any scoring metric to balance the accuracy of a structure keeping in view the structure complexity and (b) the computational tractability of any scoring function (metric). Bayesian information criterion (BIC) (Schwarz, 1978), Bayesian Dirichlet equivalence uniform (BDeu) (Buntine, 1991), Akaike information criterion (AIC) (Akaike, 1974), entropy and minimum description length (MDL) (Lam & Bacchus, 1994; Suzuki, 1996), and factorized conditional log-likelihood (fCLL) (Carvalho *et al*., 2011) have been reported to satisfy these characteristics.

Among these scoring functions, BIC, AIC, BDeu and MDL are based on log-likelihood (LL) as given below:

$$LL(G \mid D) = \sum_{i=1}^{n} \sum_{j=1}^{q_i} \sum_{k=1}^{r_i} N_{ijk} \log\left(\frac{N_{ijk}}{N_{ij}}\right) \quad ...(01)$$

Where G denotes directed acyclic graph given the dataset D. Other three counters include n, q_i and r_i representing the number of cases, the number of distinct states of a feature variable and the number of distinct states of a parent of an i^{th} feature variable. The log-likelihood tends to increase its value as the number of features increases. The phenomenon occurs because the additions of every edge are prone to pay contributions to the resultant log-likelihood of the final structure. This process can be controlled considerably by means of introducing some penalty factor or otherwise restricting the number of parents for every node in the graph.

AIC and BIC are usually applied under the hypothesis that regression orders *k* and *i* are identical. This assumption brings extra computation and also yields erroneous estimation in theoretical information measures in structure learning (Yang *et al.*, 2013). Yang & Lee (2012) demonstrated the linear impact of improvement in model quality within the scope of exercising BIC function score in K2 (Cooper & Herskovits, 1992). However, it is arguable that there must be an intelligent heuristic to sharply extrapolate the optimized size of the training data. We are of the view that an optimized solution can be achieved by exploiting various intelligent algorithms for tree and graph.

METHODS AND MATERIALS

In the previous section a brief notion of the decomposability of various scoring measures into a frequency counting problem in structure learning was given. This frequency counting problem thus defined leads to a deficiency in correctly identifying discriminative approaches in defining a sink node correctly. An improved measure of approximation based on joint and marginal probability is proposed while establishing a hypothesis such that

***Hypothesis H_1*:** I2S is a tractable approximation to correctly identify the topology between a pair of nodes in a DAG for structure learning.

It details out the relationship between two features such that the states of a dependent feature can be explained as a result of the states of the independent feature. It is essential to point out the following two assumptions for defining I2S mathematically. These include the discrete nature of the dataset features. The second assumption is that each case of the dataset holds an independent probabilistic nature. I2S can be expressed by means of definition 1 and 2 as given below:

Definition 1: Given two features F_1 and F_2, I2S can be expressed mathematically;

$$I2S_{F1,F2} = \sum_{j=1}^{n} [\nabla] \quad ...(02)$$

$$\nabla = \sum_{i=1}^{m} \left[\overline{\overline{[^{n}_{j=1}(CP_{ij})]}} - \overline{[^{n}_{j=1}(CP_{ij})]} \right] \times \frac{m}{m-1} \times MP_i \quad ...(03)$$

Conditional probability (CP) is a function of joint probability (JP) and marginal probability (MP) as shown by the equation below;

$$CP_{ij} = f(JP_{ij}, MP_j) \quad \therefore 0 \le i \le m \wedge 0 \le j \le n \quad ...(04)$$

The terms m and n point out the vector length of the 1st feature F1 and the 2nd feature F2, respectively in equation 3. There are four terms involved in the mathematical equation of I2S. The value $\overline{\overline{[^{n}_{j=1}(CP_{ij})]}}$

indicates the maximum CP among all of the states of the 2nd feature. The term $\overline{[^{n}_{j=1}(CP_{ij})]}$ shows the average CP of all of the states of the second feature. The factor m / (m-1) and MP_i are used for scaling and normalizing the factors by which the final value of I2S always pulsates between 0 and 1. In the forthcoming section, the results of various feature selection techniques will be presented as compared to this technique based on the proposed measure I2S.

***Definition 2*:** Given a directed acyclic graph (DAG), I2S is sensitive to the order of sink and its parent node. A swap will change the value of I2S such that $I2S(A,B) <> I2S(B,A)$. This characteristic is most important to correctly identify the true order of the two nodes in structure learning for decision making.

***I2S Network*:** It has been reported that the greedy approach is more popular in the application of building and learning belief networks (Carvalho *et al.*, 2011). Moreover, it has been described that K2 (Cooper & Herskovits, 1992) is one of the most optimized techniques for searching algorithms in Bayesian networks. In K2 algorithm, ordering of the features is known *a prior*, which helps in selection of the most suitable set of parents for each feature. Its input parameters are a set of nodes sorted topologically. Every node in this set is scanned, while the previous nodes are added repeatedly until the resulting score given by the joint probability of the data and the network structure is not incremented. Some notation in the light of well known and relevant concepts of discrete belief networks were introduced and these concepts were formulated into a structure learner devised on the basis of I2S. I2S is a measure defined to measure the dependency (explanation) of one feature on another feature. It is a direct measurement of cardinal relationship in a way that if any distinct value of feature 2 is addressed by only a single value of feature 1, then this will increase the value of I2S where I2S is normalized between 0 to 1. It is described formally such as Î: I2S (F_1 F_2). The notation Î will be useful in defining the value of I2S from the 1st feature (F_1) to the 2nd feature (F_2). For a dataset D, a pairwise matrix of Î can be defined;

$$M = \hat{\mathbf{I}}_{ij} \forall 1 \le (i,j) \le n \quad ...(05)$$

This matrix M is useful in the development of structure learning.

I2S Network Classifiers: An ordered list of the features was developed using I2S. Let M be a matrix in which each element corresponds to the measurement of I2S from the i^{th} feature to j^{th} feature. Let $\acute{M} \leftarrow \Phi(M)$ be defined as a list of sorted matrix where sorting criteria is defined by

$$R_i = \sum_{j=1}^{n} \hat{I}_{ij} \quad ...(06)$$

It results in an ordered list known as $\acute{M} = \{\overrightarrow{X_1}, \overrightarrow{X_2}, \ldots \ldots \overrightarrow{X_n}\}$. An I2S based network classifier is a network over X = (X_1, X_2, X_3, …X_n ,C) where feature C is considered as a class, hence the goal is to classify the instances (X_1, X_2, X_3, … , X_n) in terms of distinct states of the class. Usually, in the literature, it is common to restrict towards augmented naive Bayes classifier for the sake of computational efficiency (Carvalho *et al*., 2011); where class feature is placed at the top of the graph with null parents. This relaxation is based on the assumption that the goal is to retrieve the best possible structure, which truly represents the underlying dataset. All of the query variables, which have a parent node within a DAG must have various instances of unique states formally defined as $C \in \Pi x_1$, where C represents the unique values of the parent(s) feature. We shall introduce notations related to non-augmented naïve Bayes models. Let the i^{th} parent variable of any feature possess distinct values denoted as m_{ij}, where j is the number of unique values, which i^{th} parent holds. Hence the possible number of configurations of the parent set of any feature can be described as;

$$\prod_{j}^{n} X_j \in (\prod_{i}^{n} X_i \setminus X_j) \quad ...(07)$$

This description is useful in defining the function of conditional probability table (CPT) such as:

$$CPT(f_i) <- \prod_{j}^{n} X_j \in (\prod_{i}^{n} X_i \setminus X_j) \quad ...(08)$$

The generation of CPT turns the network into a classifier. A given instance of data can be tested against this conditional probability table for its inference or induction.

RESULTS

This section will present the results with their empirical validation in detail. The performance of the proposed measure used in introduced classifiers is measured by accuracy, which is a function of true positive (TP) rate and false positive (FP) rate. It is formally defined as;

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad ...(09)$$

Experimentation was performed on 29 datasets obtained from UCI (Blake & Merz, 1998) and was preprocessed into weka (Hall *et al.*, 2009) arff file format. No further preprocessing was done on these datasets except on

five datasets marked by (*) (Table 1), in which the class feature was placed as the last attribute (this is a mandatory requirement by weka). The flags dataset was a class-less dataset, so the feature 'religion' was fixed as its class attribute. All of these datasets contain nominal, continuous and discrete features while some datasets also contain missing cases, which were ignored by default in weka. It is evident from Table 1 that the dataset is versatile in the number of classes, cases and attribute count so that no question of bias can be raised.

Table 1: Statistical information about dataset used in this study

Dataset	Cases	Attribute count	Class
Arrhythmia	452	280	13
Audiology	226	70	24
balance-scale	625	5	3
breast-cancer	286	10	2
breast-w	699	10	2
bridges_version1	107	13	6
bridges_version2	107	13	6
credit-a	690	16	2
Dermatology	366	35	6
Diabetes	768	9	2
Ecoli	336	8	8
flags*	194	30	8
Glass	214	10	6
Haberman	306	4	2
heart-h	294	14	2
hepatitis	155	20	2
liver-disorders	345	7	2
lung-cancer	32	57	2
mfeat-morphological	2000	7	10
molecular-biology promoters	106	59	4
postoperative-patient-data	90	9	3
primary-tumor	339	18	21
shuttle-landing-control*	15	7	2
solar-flare_1*	323	13	6
solar-flare_2*	1066	13	6
soybean	683	36	19
spect_test*	187	23	2
Tae	151	6	3
Zoo	101	17	7

Figure 1, which is a stacked cylindrical graph indicates the comparison of result accuracy for six scoring functions and introduced measures. Each cylinder is shown in three colours. The blue colour indicates the percentage of datasets in which the performance of I2S was significantly better than the other scoring function. The red colour indicates the number of datasets where the proposed measure neither delivers better nor demonstrates poor accuracy in classification. The green colour indicates the number of datasets in which I2S failed to yield better results. A careful examination of Figure 1 shows that the accuracy of I2S was comparably higher in comparison to AIC and entropy, where I2S delivers improved accuracy over 22 and 21 datasets while it does not give better results over 3 and 5 datasets, respectively. The recently introduced scoring function measure fCLL gives comparatively better accuracy in comparison to the other five scoring functions when competing with I2S.

Apart from the results shown in Figure 1, one may argue that achieving accuracy may not be so impressive; whereas the percentage improvement in accuracy is more compelling. This motivates the introduction of results from another perspective shown in Figure 2, which indicates the percentage of average improvement of accuracy achieved by using the I2S classifier in the K2 searching algorithm. In the case of the entropy measure, the average increase in accuracy was observed as more than 7.5 % while it was 1.19 % in comparison to BDeu.

To roughly characterize the computational complexity of the proposed scoring measure, it was noted that the time complexity of 12S was more or less equivalent to that of BDeu and BIC. However, the time complexity of entropy was slightly better than I2S. Moreover, the time complexity for AIC and MDL was significantly better than I2S in many of the datasets.

CONCLUSION AND FUTURE WORK

In classification, structure prediction from Bayesian inference models is a common practice for the purpose of retrieving hidden rules from masses of data. This process broadly consists of two steps. The first step deals with the construction of the best suitable structure from the data and the second part with the inference from this structure. This study was focused on the first part, which involved the construction of the most suitable network structure.

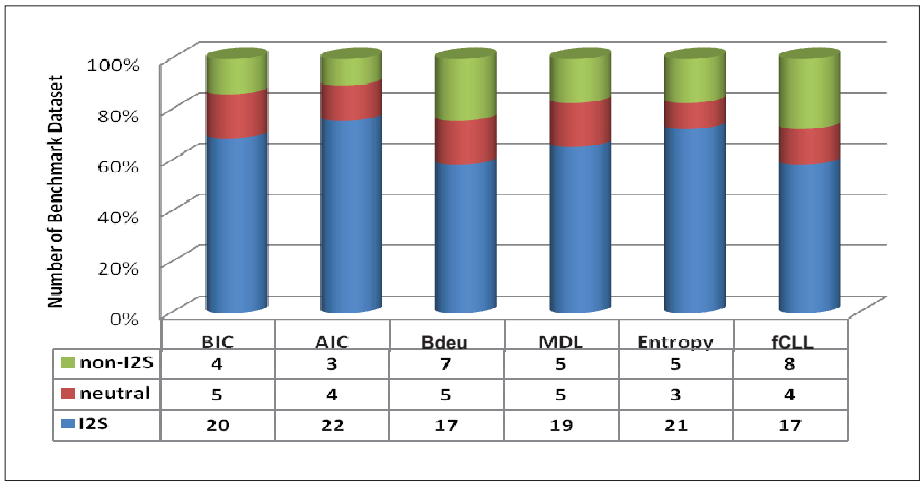

Figure 1: Comparison of accuracy of I2S vs other scoring functions

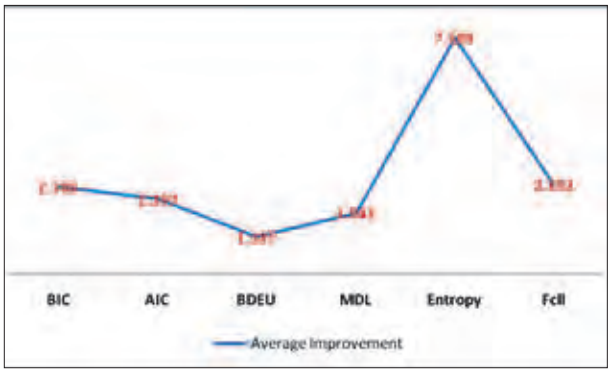

Figure 2: Average improvement (%) of I2S vs other scoring functions

The core part in the design of a BBN classifier is to introduce a discriminant function within the vector space of attributes through utilization of *a priori* knowledge. The effectiveness of the Bayesian belief network using greedy heuristics like the K2 searching mechanism has earned it an excellent place in the domain of classification systems. Arguments were presented about various scoring functions including BDeu, AIC, entropy, BIC, MDL and a recently introduced fCLL on the ground of overfitting while introducing a new dependency measure in the domain of structure learning. Theoretically, application of mutual information in structure learning is not a novel idea as it was introduced some six decades ago (Chow & Liu, 1968; Pearl, 1988). In this study a novel decomposable scoring function was introduced for the task of structure learning. The introduced measure, known as integration to segregation is characterized by the mutual dependence approximated by marginal and joint probability. The novel measure is particularly designed for discriminative learning because it is decomposable and score-equivalent with the capability of permitting efficient estimation of structure learning. The accuracy merit of I2S is evaluated and compared to the common state of-the-art scoring measures given a reasonable size of benchmark datasets obtained from the UCI repository and preprocessed in weka. I2S performed better than generatively-trained Bayesian network classifiers using K2 searching algorithm and numerous scoring functions. The proposed measure is expected to generate a realistic network, which is likely to tally with the practical thinking of field experts in the domain of knowledge. Although the asymptotic complexity of the proposed measure is almost of the same order as the conventional BIC and BDeu scoring metrics, it is still poor in computational complexity as compared to MDL in particular.

Acknowledgement

We are greatly thankful to anonymous reviewers who suggested numerous insightful comments during the revision of this article.

REFERENCES

1. Akaike H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control* **19**:716 − 723. DOI: http://dx.doi.org/10.1109/TAC.1974.1100705
2. Aldrich J. (1995). Correlations genuine and spurious in Pearson and Yule. *Statistical Science* **10** (4): 364 − 376.
3. Bagdonavicius V., Kruopis J. & Nikulin M.S. (2011). *Non-parametric Tests for Complete Data*. ISTE & Wiley: London & Hoboken.
4. Blake C. & Merz C. (1998). UCI repository of machine learning databases. Available at *http://www.ics.uci.edu/~mlearn/, MLRepository.html*, Accessed December 2012.
5. Buntine W.L. (1991). Theory refinement on Bayesian networks, *Proceedings of the 7th Conference on Uncertainity in Artificial Intelligence*, Los Angeles, USA, 13-15 July, pp. 52 − 60.
6. Carvalho A.M., Roos T.T., Oliveira A.L. & Myllymäki P. (2011). Discriminative learning of Bayesian networks via factorized conditional log-likelihood. *Journal of Machine Learning Research* **12**: 2181 − 2210.
7. Chen X-W., Anantha G. & Lin X. (2008). Improving Bayesian network structure learning with mutual information-based node ordering in the K2 algorithm. *IEEE Transactions on Knowledge and Data Engineering* **20** (5): 628 − 640.
8. Chow C.K. & Liu C.N. (1968). Approximating discrete probability distributions with dependence trees. *IEEE Transactions on Information Theory* **14**: 462 − 467. DOI: http://dx.doi.org/10.1109/TIT.1968.1054142
9. Cooper G.F. & Herskovits E. (1992). A bayesian method for the induction of probabilistic networks from data. *Machine Learning* **9**: 309 − 347. DOI: http://dx.doi.org/10.1007/BF00994110
10. Corder G.W. & Foreman D.I. (2009). *Nonparametric Statistics for Non-Statisticians: A Step-by-Step Approach*, pp.134. Wiley, New Jersy, USA. DOI: ht tp://dx.doi.org/10.1002/9781118165881
11. Etminani K., Naghibzadeh M. & Razavi A.R. (2010). Globally optimal structure learning of bayesian networks from data (eds. K. Diamantaras, W. Duch & L.S. Iliadis), *20th International Conference on Artificial Neural Networks*, Thessaloniki, Greece, September 2010, pp. 101 − 106.
12. Gibbons D. J. & Subhabrata C. (2003). *Nonparametric Statistical Inference*, 4th edition, pp.399. CRC Press, UK.
13. Grimmett G. & Stirzaker D. (2001). *Probability and Random Processes*, p.27. Oxford University Press, Oxford, UK.
14. Hall M., Frank E., Holmes G., Pfahringer B., Reutemann P. & Witten I.H. (2009). The Weka data mining software: an update. *SIGKDD Explorations* **11**: 10 − 18. DOI: http://dx.doi.org/10.1145/1656274.1656278
15. Jensen F.V. & Nielsen T.D. (2007). Bayesian networks and decision graphs. *Information Science and Statistics*, pp.242. Springer, New York, USA. DOI: http://dx.doi.org/10.1007/978-0-387-68282-2_2
16. Lam W. & Bacchus F. (1994). Learning Bayesian belief networks: an approach based on the MDL principle. *Computational Intelligence* **10**:269 − 294. DOI: http://dx.doi.org/10.1111/j.1467-8640.1994.tb00166.x
17. Pearl J. (1988). *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann, Burlington, Massachusetts, USA.
18. Schwarz G.E. (1978). Estimating the dimension of a model. *Annals of Statistics* **6** (2): 461 − 464. DOI: http://dx.doi.org/10.1214/aos/1176344136
19. Suzuki J. (1996). Learning Bayesian belief networks based on the MDL principle: an efficient algorithm using the branch and bound technique, *Proceedings of the International Conference on Machine Learning*, Bally, Italy.
20. Wasserman L. (2007). *All of Nonparametric Statistics*. Springer Science, USA.
21. Yang C., Le Bouquin Jeannes R., Bellanger J. & Shu H. (2013). A new strategy for model order identification and its application to transfer entropy for EEG signals analysis. *IEEE Transactions on Biomedical Engineering* **99**: 1−1.
22. Yang L. & Lee J. (2012). Bayesian belief network-based approach for diagnostics and prognostics of semiconductor manufacturing systems. *Robotics and Computer-Integrated Manufacturing* **28**: 66 − 74. DOI: http://dx.doi.org/10.1016/j.rcim.2011.06.007