An Information Theoretic Scoring Function in Belief Network

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

Abstract: *We proposed a novel measure of mutual information known as Integration to Segregation (I2S) explaining the relationship between two features. We investigated its nontrivial characteristics while comparing its performance in terms of class imbalance measures. We have shown that I2S possesses characteristics useful in identifying sink and source (parent) in a conventional directed acyclic graph in structure learning technique such as Bayesian Belief Network. We empirically indicated that identifying sink and its parent using conventional scoring function is not much impressive in maximizing discriminant function because it is unable to identify best topology. However, I2S is capable of significantly maximizing discriminant function with the potential of identifying the network topology in structure learning.*

Keywords: *Mutual dependence, information theory, structure learning, scoring function.*

Received September 9, 2012; accepted March 21, 2013; published online February 26, 2014

1. Introduction

Measurement of the dependence relationship and correlation among features in a given dataset is an interesting and fundamental problem in the domain of classification. Numerous pair wise measures have been proposed describing a sensible relationship in general or in specific context. The detail of these measures can be obtained from literature [10, 12, 26]. It was reported that correlation and dependence both intrinsically are quite different phenomenon. The roots of many such measures rest in information theory, whereas mutual information was first introduced by Shannon in domain of digital communication. It was described that the mutual information among two random features is symmetric. This property lay down the foundation of its capability to be used in capacity of dependence measure. We in this study have interrogated this fact that in the domain of structure learning classifiers, the symmetric characteristic in fact does not imply correct meaning.

Measuring dependence rests at the heart of various statistical problems. Classification is one kind of such problem for which measurement of dependence plays an important role. Regardless of wide application of correlation in various domains of scientific knowledge, a careful examination of correlation measures in general reveal two issues related to problem solution towards structure learning. The first issue is related to its inability of explaining nonlinear structure between the random features. It was elaborated that two independent features are certainly uncorrelated but, being uncorrelated does not mean they are necessarily independent to each other [13]. The second issue is it's in capability of providing limited information around the underlying true dependence nature [13]. This leads to arise a dictum that "correlation is unable to imply causation" means that correlation is not ideally well suited in classification problem for sake of delivering causal relationship between the features [3].

When we talk about structure learning, then Bayesian Belief Network (BBN) can't be brushed aside. Since last two decades, the Bayesian belief network (also, known as BBN) has inspired a lot of communities dealing in knowledge management and pattern classification [1, 11, 20, 24]. The BBN is a highly symbolic formalism probabilistic model for knowledge representation. A BBN is a Directed Acyclic Graph (DAG) representing a set of conditional probability distribution for each node of the DAG whereas each arc between two nodes represent direction of inference or induction. A node (child) which is directly pointed to by another node (parent) has inference from its parent node(s), while the parent node receives induction from the child node in terms of probabilistic distribution. These concepts of inference and induction are helpful in formulation of BBN classifier. The topology or ordering of the child parent node is important in the evaluation of class imbalance measures for a BBN classifier. We in this study have shown that Integration to Segregation (I2S) has the ability to correctly identify the true order between two nodes to place their role for being parent or child to each other by virtue of its peculiar characteristic of being asymmetric in nature. Many correlation measures are not only symmetric but also, linear in nature. Pearson correlation coefficient is a notable example which is quite famous and has been used extensively because of its low computational cost and

ease of estimation. However, in most of the cases, the dataset is not necessarily linear in nature. Our proposed measure I2S is applicable to linear as well as nonlinear data sets.

The aim of this study is introduce a novel dependency measure which is applicable to be used in capacity of a scoring functions used by learning algorithms paradigm in the context of classification, namely, for learning the BBN classifier. We make an attempt to empirically evaluate the merits of each score as compared to our proposed measure I2S by means of a experimental study over benchmark datasets. These scores include well renowned scoring function including Bayesian Information Criterion (BIC), Akaike Information Criterion (AIC), Bayesian Dirichlet (BDeu) with likelihood equivalence and a uniform joint distribution, Entropy and Minimum Description Length (MDL). We shall detail out their mathematical description in the next section. Recently Carvalho *et al.* [7] introduced factorized Conditional Log-Likelihood (fCLL) which is an approximation of the conditional log-likelihood criterion. It was a non parametric measure, which was claimed empirically to give significantly better results as compared to other measures in K2 and TAN algorithms. Such impressive scoring function motivates us to analyze its performance over correctly judging the true topology, the result of which is shown in our result section.

The organization of the remaining paper is articulated as below: In section 2, some background of the proposed measure is discussed. In section 3, mathematical theory of the proposed dependence measure is presented. We have devoted last two sections for empirical validation of I2S followed by discussion in detail.

2. Related Work

In last two decades, a lot of work has been reported in improvements of BBN. Application of BBN in various domain of interest has been highlighted [4, 25]. Cheng *et al.* [8] categorized BBN classifiers into two groups. One is scoring based method whereas various scoring criteria were introduced such as: Bayesian scoring method [9, 15], entropy based method [16] and minimum description length method [23]. The other group focuses on analysis of dependence relationships among features under Conditional Independence (CI) test. The algorithms described in [8, 22, 27] belong to this category. Meilaˇ and Jaakkola [21] elaborated Bayesian structure learning in trees in polynomial time. In structure learning, there are two dimensions for the application of features dependence. The first criterion is selection of useful features under the assumption of maximal statistical dependence criteria [21]. The second category is the appropriate selection of features in the designing of markov blanket for a node as a potential candidate class [8]. In the first category, poor features can be eliminated; however it does not guarantee a more suitable and accurate set of markov blanket for a class node. Statistical analysis before the application of classifier on dataset is recommended [26]. This motivates us to adopt the second category in which careful selection of markov blanket nodes is more important.

BIC score which was originally introduced by Schwarz [21] was framed over belief network with hidden features. Jensen *et al.* [17] discussed two important characteristics for scoring function used in the belief network. The first characteristic is the ability of any score to balance the accuracy of a structure in context of structure complexity. The second characteristic is its computational tractability. BIC is believed to satisfy both of the mentioned characteristics. BIC is formally defined as:

$$BIC(S / D) = \log_2 P(D / \theta_S^n, S) - \frac{size(S)}{2} \log_2 (N) \quad (1)$$

Where $\hat{\theta}$ is an estimate of the maximum likelihood parameters for the underlying structure. Jensen *et al.* [17] discussed that in case of completion of the database, BIC is reducible into problem of determination of frequency counting such as:

$$BIC(S \mid D) = \sum_{i=1}^{n} \sum_{j=1}^{q_i} \sum_{k=1}^{r_i} N_{ijk} \log_2 \left(\frac{N_{ijk}}{N_{ij}}\right) - \frac{\log_2 N}{2} \sum_{i=1}^{n} q_i (r_i - 1) \quad (2)$$

Where N_{ijk} indicates the counts of dataset cases with node X_i in its k^{th} configuration and $pa(Xi)$ in j^{th} configuration, q_i denotes the number of configurations over the parents for node X_i in space S and r_i indicates the states of node X_i. Another scoring measure which depends only on equivalent sample size N' is BDeu [6]. Carvalho *et al.* [7] has decomposed it into mathematical form:

$$BDeu(B,T) = \log(P(B)) + \Delta \quad (3)$$

$$\Delta = \sum_{i=1}^{n} \sum_{j=1}^{q_i} \left(\log \left(\frac{\Gamma\left(\frac{N'}{q_i}\right)}{\Gamma\left(N_{ij} + \frac{N'}{q_i}\right)} \right) + \sum_{k=1}^{r_i} \log \left(\frac{\Gamma\left(N_{ijk} + \frac{N'}{r_i q_i}\right)}{\Gamma\left(\frac{N'}{r_i q_i}\right)} \right) \right) \quad (4)$$

MDL [18, 23] is mostly suitable to complex Bayesian network. Mathematical formulation is composed of explanation of Log Likelihood (LL) as following:

$$LL(B \mid T) = \sum_{i=1}^{n} \sum_{j=1}^{q_i} \sum_{k=1}^{r_i} N_{ijk} \log \left(\frac{N_{ijk}}{N_{ij}} \right) \quad (5)$$

The value of LL is used in obtaining the value of MDL as below:

$$MDL(B \mid T) = LL(B \mid T) - (1/2) \log(N) \mid B \mid \quad (6)$$

|B| denotes the length of network which is achieved in terms of frequency calculation of a given feature's possible states and its parent's state combination with feature as following:

$$| B | = \sum_{i=1}^{n} \left(r_i - 1 \right) q_i \tag{7}$$

Akaike Information Criterion (AIC) originally defined by Akaike [5] is defined mathematically:

$$AIC = -2 \times ln(likelihood) + 2 \times K \tag{8}$$

Where K denotes the number of parameters in the given model. However, in belief network application, its mathematical equation has been transformed into:

$$AIC(B \mid T) = LL(B \mid T) - | B | \tag{9}$$

Recently fCLL was introduced which is an approximation of the conditional log-likelihood criterion [7]. Its decomposability over the network structure is defined as below:

$$\begin{aligned} \hat{f}\, CLL(G \mid D) &= (\alpha + \beta) \hat{LL}(B \mid D) \\ -\beta\lambda \sum_{i=1}^{n} \sum_{j=1}^{q_i^*} \sum_{k=1}^{r_i} \sum_{c=0}^{1} N_{ijkc} \left(log\left(\frac{N_{ijkc}}{N_{ij*c}} \right) - log\left(\frac{N_{ijc}}{N_{ij*}} \right) \right) \\ -\beta\lambda \sum_{c=0}^{1} N_c \, log\left(\frac{N_c}{N} \right) - \beta N \rho \end{aligned} \tag{10}$$

The authors empirically showed that their measure of scoring function is not only decomposable to Bayesian network structure learning but it also, can given better results over various other conventional measures. We have added a comparison of its performance to I2S in result section.

3. Development of Scoring Measure

We in previous section deliver a brief notion of reducing various scoring function into a frequency counting problem in terms of structure learning. This frequency counting problem defined in Equation 2 leads to a shortcoming of correctly identifying discriminative approaches in defining sink node correctly. We come up with an improved measure of approximation while establishing a hypothesis such that

3.1. Hypothesis H_1

I2S is an amenable approximation to correctly identify a sink between a pair of nodes in a DAG in structure learning.

Reasoning for our hypotheses is based on obtaining structural inference from dataset. The primary goal of this research is to design and develop a robust mutual dependence measure which can maximize structure prediction technique applicable to generate a directed acyclic graph from the dataset. The novel measure possesses crucial properties which are useful in structure learning procedure technique. The underlying hypothesis describes the relationship between two features such that dependent feature can be explained as a consequence of the action of the independent features. It is useful to draw attention on the following two underlying assumptions for description of I2S. These include the discrete nature of the dataset with no missing values. The second assumption is that each case of the dataset enjoys independent probabilistic nature. It is convenient to define the I2S formally by means of some definitions and lemma.

3.2. Definition 1

Given two features A and B, f is a relation on $A.B$ such that $f: \phi(Y) \rightarrow X$ where $domain(Y)$ = {A, B} and for every $a \in A$, there exist precisely one $b \in B$ while $(a, b) \in X$. ϕ is a real-valued function over a domain of finite variables Y whereas members of domain (Y) need not to be linked in conditional probability. Every element of X is comprised of data from each distinct state of B determined by unique outcome states of A.

3.3. Definition 2

Let X be a matrix as defined in definition 1 then value of I2S is defined as:

$$Y_{ij} = X_{ij} / B_j \quad \therefore 0 \le i \le m \wedge 0 \le j \le n \tag{11}$$

$$I2S = \sum_{j=1}^{n} ([Max_{j=1}^{n}(Y_{ij}) - Average_{j=1}^{n}(Y_{ij})] \times [m / (m-1)] \times B_j) \tag{12}$$

3.4. Definition 3

Let feature set A and B comprise of m and n number of distinct states. Let $X \rightarrow A \bullet B$ be a matrix as defined in definition 1 such that $\sum_{i}^{m} \sum_{j}^{n} P_{ij} = 1$. It is clear that each column of matrix X is determined by corresponding single outcome state of feature A with combination of all items of feature B. Mathematically we can write it:

$$\sum_{i=1}^{n} X_{i1} = P(a_1) \tag{13}$$

$$\sum_{i=1}^{n} X_{i2} = P(a_2) \tag{14}$$

$$\sum_{i=1}^{n} X_{im} = P(a_m) \tag{15}$$

It can also, be concluded that each row of matrix X is determined by corresponding single state of feature B with combination of all items of feature A. Mathematically we can write it:

$$\sum_{j=1}^{m} X_{1j} = P(b_1) \tag{16}$$

$$\sum_{j=1}^{m} X_{2j} = P(b_2) \tag{17}$$

$$\sum_{j=1}^{m} X_{nj} = P(b_n) \tag{18}$$

3.5. Lemma 1

Let X be a matrix as defined in definition 1. If every element of matrix X holds same probability then this set can be termed a fully integrated matrix. In other words, a matrix X is fully integrated in which all of the information is uniformly distributed over all of its

elements such that: $P(x_{ij})=P(x_{ji})$. In this case, value of I2S for matrix X will be zero.

Proof: Let A and B be two features in a dataset. A and B comprise of m and n distinct state outcome respectively such that:

$$P(A_1) + P(A_2) + ... + P(A_m) = \sum_{i=1}^{m} P(A_i) \quad (19)$$

$$P(B_1) + P(B_2) ..... + P(B_n) = \sum_{j=1}^{n} P(B_j) \quad (20)$$

Let feature set A and B comprise of m and n number of distinct states. Let $X \rightarrow A \bullet B$ such that each element of X corresponds to $a \in A$ and $b \in B$ such that:

$$\sum_{i}^{m} \sum_{j}^{n} P_{ij} = 1 \quad (21)$$

Definition 3 gives the orientation of the matrix in a way to deduce the relationship between features A, B to matrix X. It is clear that when definition 2 is applied on the matrix X, then it will decrease each value by same factor in each row and column monotonically. By definition 2, Max and Average of each row is also, same in this case. By definition, $I2S_i=0$ for each row in matrix. From this explanation we can conclude that:

$$\sum_{j=1}^{n} I2S_{ij} = 0 \quad (22)$$

Hence, it is proved that matrix is fully integrated.

3.6. Lemma 2

If we distribute probability of each element in matrix X in a non uniform fashion then its segregation will increase at expense of integration.

Proof: Let A and B be two features in a dataset. A and B comprise of m and n distinct state outcome respectively such that:

$$P(A_1) + P(A_2) ..... + P(A_m) = \sum_{i=1}^{m} P(A_i) \quad (23)$$

$$P(B_1) + P(B_2) ..... + P(B_n) = \sum_{j=1}^{n} P(B_j) \quad (24)$$

Let $X \rightarrow A \bullet B$ given that each element of X corresponds to $a \in A$ and $b \in B$ such that: $\sum_{i}^{m} \sum_{j}^{n} P_{ij} = 1$.

Now, let us choose a single element P_{kl} where $0 <= K <= m$ and $0 <= l <= n$. We increase its probability at the expense of decreasing probability of other elements in a random fashion but with condition: $\sum_{i}^{m} \sum_{j}^{n} P_{ij} = 1$.

Definition 3 gives the orientation of the matrix in a way to deduce the relationship between features A, B to matrix X. As we applied definition 2 on the matrix X, it will decrease each value by same factor in each row and column monotonically. By definition 2, Max and Average of each row is not same in this case. Hence: by definition, $I2S_{ij}>0$ for at least in the row containing the element P_{kl}. From this explanation we can conclude that: $\sum_{j=1}^{n} I2S_{ij} > 0$.

Hence, it is proved that matrix has lost its integration at the expense of increase in its segregation.

3.7. Lemma 3

Matrix X is said to be fully segregated with value of I2S approaching to 1 if each distinct state of feature B is selected by not more than one distinct state of feature A.

Proof: Let A and B be two features in a dataset. A and B comprise of m and n distinct outcomes respectively as shown in equations 23 and 24. Let feature set A and B comprise of m and n number of distinct states. Let $X \rightarrow A \times B$ such that each element of X corresponds to $a \in A \wedge b \in B$ such that: $\sum_{i}^{m} \sum_{j}^{n} P_{ij} = 1$.

Now, let us choose a single element P_{kL} where $0 <= K <= m$ and $0 <= L <= n$. We increase its probability at the expense of decreasing probability of other elements in a random fashion but with condition: $\sum_{i}^{m} \sum_{j}^{n} P_{ij} = 1$.

Definition 3 gives the orientation of the matrix in a way to deduce the relationship between features A, B to matrix X. We repeat this process for each row, till matrix X is transposed into a X' such that it contains approximately all of its probability distribution in m number of elements. As we applied definition 2 on the matrix X, then it will decrease each value by same factor in each row and column monotonically. By definition 2, Maximum and Average of each row is not same in this case. By application of definition 2 we will get $\sum_{j=1}^{n} I2S_{ij} \gg 0$ for complete matrix. From this explanation we can conclude that the matrix is fully segregated. In other words, we can say that if we distribute probability of each element in matrix X in such a way that one element in each row of matrix X contains maximum probability out of the total sum of the probability of that row, and then such matrix is fully segregated with value approaching to 1.

3.8. Definition 4

Given a DAG, I2S is sensitive to 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 two nodes in structure learning for decision making.

3.9. Definition 5

The measure I2S is minimum if all the input probability measures become equiprobable. On the other hand the measure tends to maximized if there is maximum difference found among probability distribution between combined probability distribution such that f: Ð(Θ) where Ð is a function to calculate the

maximum difference between probability distribution parameters Θ in a cross product of both of the features sink and its corresponding parent in a DAG.

3.10. Lemma 4

Given two nodes in BBN, I2S can detect the best topology of two nodes.

Proof: Let *A* and *B* be two features such that *A* holds deterministic function *B* (sink). Feature *A* contains m distinct states and *B* contains n distinct states.

If we are to get inference of *A* on *B* such that (*B* | *A*) then a conditional probability table *C* is formulated such as $\{C_{ij} / (i <= m)$ and $(j <= n)\}$. The objective of the naïve Bayes classifier is to:

$$Maximize \begin{bmatrix} f_1 : C_{11}, C_{12}, \ldots\ldots C_{1i} \\ f_2 : C_{21}, C_{22}, \ldots\ldots C_{2i} \\ f_3 : C_{i1}, C_{i2}, \ldots\ldots C_{ii} \end{bmatrix} \quad (25)$$

There are two possible orientations for the topology of two nodes. $A \rightarrow B$ or $B \rightarrow A$. We apply integration to segregation measure on both of these topologies such that: $I2S_1$: $A \rightarrow B$ and $I2S_2$: $B \rightarrow A$.
As *I2S* is an asymmetric measure therefore:

$$\begin{bmatrix} I2S_2 - I2S_1 > 0 \\ I2S_2 - I2S_1 = 0 \\ I2S_2 - I2S_1 < 0 \end{bmatrix} \quad (26)$$

Every function in Equation 1 will be increased with the increase in the difference between its corresponding variables. The same is true for *I2S* when each *I2S* will be increased as its underlying features are getting less segregated with increase in integration. Hence, it is proved that an increase in *I2S* is analogous to maximization of probabilistic function defined in Equation 1. Hence we can deduce *I2S* can be used to define quite correct topology of two nodes in BBN.

Algorithm 1: I2S value out of two features F_1 and F_2.

Input: F_1, F_2
Output: I2S
Function [I2S] = Find_I2S(F_1, F_2)
Step 1. M = GPM(F1, F2)
Step 2. $M = \frac{M}{\oplus M}$
Step 3. $B = \overleftarrow{\oplus c} M$
Step 4. sz ← |M(:,1)|
Step 5. V ← zeros(sz,1)
Step 6. For each row r ε M
6.1 M(r,:) ← M(r,:) / B(r,1)
*6.2 V(r,1) ← I2SE(M(r,:)) * B(r,1)*
6.3 r ← r +1
6.4 if r > sz
6.5 exit for
6.6. end if
Step 7. end while
Step 8. $I2S = \overset{\oplus}{V}$
Step 9. End

It is convenient to decompose the whole of the computational methodology into three functions to elaborate it in a sensible flow. Algorithm 1 is prime part of the computational methodology of calculation of value of I2S. It begins with two input parameters F_1 and F_2 giving final value of I2S between both of these features. In line 2, the function has called a function GPM which is responsible for generation of pair wise matrix out of two features. This function (GPM) has been shown in algorithm 2 where each of the features is first subjected to calculate the unique values involved in the feature. A Matrix M is born given the dimension of length of unique values of feature 1 and 2 as number of rows and columns of matrix M respectively. These steps are pointed out from step 2 to step 5 in which unique values are calculated *U1, U2* and then length of the unique values *L1, L2* are calculated. For the convenience of understanding of the procedure outlined in this function, we have made assumption that in each feature the unique values are consecutive. This can be achieved by first decoding each feature attribute of the dataset. The rest of the lines in the function (GPM) simply count the number of distinct values of both of the features in a pair wise fashion by populating the resulting matrix M. Now, we shall again continue our explanation on our prime function Find_I2S.

Algorithm 2: generate pair-wise matrix

Input: F1 , F_2
Output: Trans
Function [M] = GPM(F_1, F_2)
Step 1. $U_1 = \overset{\triangleleft}{F}_1$
Step 2. $L_1 = \overleftarrow{U}_1$
Step 3. $U_2 = \overset{\triangleleft}{F}_2$
Step 4. $L_2 = \overleftarrow{U}_2$
Step 5. M = zeros(L_1,L_2)
Step 6. L = |F_1(:,1)| // rows count in F1
Step 7. for each i ε L
Step 8. M(F_1(i,1),F_2(i,1)) = M(F_1(i,1),F_2(i,1)) + 1
Step 9. end for
Step 10. End

In line 3 this matrix is normalized such that each value ranges between 0 and 1 with total sum of all elements of matrix be only1. This step first adds up all of the values in the matrix storing it in a temporary scalar value and secondly divides each element of matrix by this scalar value. In step 4, the function extracts a vector *B* from matrix M by summing up each column of the pair wise matrix M. In the line 5, a threshold variable *sz* is introduced which is used in traversal loop from step 7 to step 14. In step 6, a vector V is created with same number of elements as that of the count of rows in the matrix M. From line 7 to line 14, a loop is launched with following activities. First

each row of the matrix M is updated by dividing corresponding element of vector B such that row number of both numerator and denominator are same.

Another function *I2SE* is called on in line 8. This function is responsible to define each of the individual elements of the final outcome vector of I2S. The underlying logic to compute the value of each element of vector V is determined by finding maximum and mean of the each row of the pair wise matrix M. A difference of maximum and minimum if multiplied by number of elements in the row. The resulting scalar value is divided by number of element in the row with degree of freedom 1.

Algorithm 3: calculate I2S elemental value

Input: Inp
Output: I2S
Function [ret] = I2SEl (Inp)
Step 1. $mx = \overleftarrow{Inp}$
Step 2. $mx = \overline{Inp}$
Step 3. $cnt = |\overset{col}{Inp}|$
*Step 4. ret = (mx - mn) * (cnt)/ (cnt-1)*
Step 5. End

At the end of loop which is marked by threshold value sz, we achieve a final vector V with same number of elements as that of the number of rows in matrix M. In line 15, a complete sum of all of the elements in the vector is obtained which is our final desirable outcome value known as I2S.

In this section, we discussed mathematical theory of the proposed measure. First we show that theoretically asymmetric property of dependence measure is more important in judgment of sink and source node. In the second part of this section, we highlighted algorithmic steps involved in the calculation of I2S. In the next section of result, we present empirical results to validate our claims made in the current and previous sections.

4. Results and Discussion

We carried out experimentation on fourteen benchmark dataset obtained from UCI [5] as shown by Figures 1 and 2. All of these dataset contain multinomial/ categorical features; hence, they were fit for our experimentation. First we explain the detail of the experiment as shown by the Tables 1 and 2. Both of these tables indicate detail of a few experiments performed on one benchmark dataset Zoo [5]. First two columns of the Table 1 indicates two features in which one is designated as sink while other is marked as class node. The last four columns of the Table 1 shows class imbalance characteristic (accuracy measures) of the experiment including Recall or True Positive Rate (TPR), False Positive Rate (FPR), Precision and F-Measure. In literature, it was termed that maximization of the scoring function guarantee the correct structure of learning. Thus, we have provided this information under the assumption that there must be a relationship of scoring function versus these class imbalance characteristics to judge how accurate a structure is defined while describing the position of both of the nodes. Table 2 indicates a gradual increase and decrease in the value of I2S versus other accuracy measures.

Table 1. Class imbalance characteristics (Zoo dataset).

S.	Sink	Class	Recall	FPR	Prec	F-M
1	1	2	0.802	0.802	0.643	0.714
	2	1	0.624	0.279	0.8	0.59
2	3	7	0.554	0.554	0.307	0.396
	7	3	0.584	0.584	0.341	0.431
3	11	17	0.406	0.36	0.179	0.248
	17	11	0.921	0.921	0.848	0.883
4	4	5	0.762	0.762	0.581	0.66
	5	4	0.545	0.458	0.559	0.549
5	6	9	0.822	0.822	0.675	0.741
	9	6	0.644	0.644	0.414	0.504
6	10	16	0.535	0.572	0.458	0.436
	16	10	0.792	0.792	0.627	0.7
7	8	12	0.832	0.832	0.692	0.755
	12	8	0.604	0.604	0.365	0.455
8	13	14	0.832	0.46	0.845	0.805
	14	13	0.406	0.28	0.194	0.261
9	15	16	0.525	0.595	0.308	0.388
	16	15	0.871	0.871	0.759	0.811
10	4	13	0.465	0.206	0.417	0.439
	13	4	0.832	0.192	0.831	0.831

There were seventeen features available in the zoo dataset for which 136 ($17\times16\rightarrow272/2$) pair-wise structures can be built. We pick only three pair results in Tables 1 and 2 randomly ten pairs in such a way that almost every feature can be shown by the Table 1. We justify the adoption of pair wise sets of features in terms of introduction of non augmented classifier in which any feature can be considered as a class while others features are treated for a selection of its corresponding potential markov blanket. Take an example from Table 1 in which serial 1 is indicating two pairs with feature number one and feature number two. Firstly, feature two was considered parent while the other feature number one was assumed as its sink or single markov blanket node. In second experiment, positions of both nodes were swapped. We analyzed accuracy measures against scoring functions as shown by Tables 1 and 2. This part of experiment for reasoning of uncertainty was performed in WEKA [14]. General parameters for classification in WEKA experimentation were simple estimator and 10 fold cross validation while K2 score type chosen is Bayes. A careful examination of Tables 1 and 2 indicates that a higher value of I2S ensures a higher precision, recall, TPR and lower FPR for example when feature 1 is selected as parent and feature 2 as sink node then Recall TPR, FPR, Precision and F-Measure values are

0.802, 0.802, 0.643 and 0.714 respectively. A swap of both of the features yields the values of 0.624,0.279,0.8 and 0.59 for Recall TPR, FPR, Precision and F-Measure. Here, the question arises how we can relate it to the scoring values presented in Table 2.

Table 2. Comparison of scoring function vs. I2S (Zoo dataset).

S.	I2S	Bayes	Bdeu	MDL	ENTROPY	AIC
1	0.604	-113.29	-115.01	-117.15	-110.23	-113.23
	0.2475	-113.51	-115.22	-117.36	-110.44	-113.44
2	0.1089	-144.88	-146.6	-147.01	-140.08	-143.08
	0.1683	-144.88	-146.59	-147	-140.08	-143.08
3	0.33	-208.07	-233.68	-233.33	-203.33	-216.33
	0.8416	-204.39	-234.06	-233.67	-203.68	-216.68
4	0.5248	-122.41	-124.13	-125.34	-118.42	-121.42
	0.2079	-122.55	-124.27	-125.48	-118.56	-121.56
5	0.6436	-119.76	-121.48	-122.46	-115.53	-118.53
	0.2871	-119.98	-121.7	-122.67	-115.75	-118.75
6	0.1287	-125.5	-127.22	-128.14	-121.21	-124.21
	0.5842	-125.3	-127.02	-127.94	-121.02	-124.02
7	0.6634	-110.88	-112.6	-114.78	-107.86	-110.86
	0.2079	-111.14	-112.86	-115.04	-108.12	-111.12
8	0.703	-199.87	-222.56	-223.45	-198.07	-209.07
	0.2871	-199.76	-219.46	-220.61	-195.22	-206.22
9	0.1287	-114.85	-116.56	-117.71	-110.79	-113.79
	0.7426	-114.46	-116.18	-117.33	-110.41	-113.41
10	0.4059	-203.84	-223.54	-224.78	-199.39	-210.39
	0.6634	-204.5	-227.18	-228.16	-202.77	-213.77

Table 2 indicates that value of I2S and all other scoring measures is also, greater in the same fashion, means a higher precision indicates a higher value of scoring function value. This trend is quite explicit and is limited to only seven set of pair of experiments. However, serial number 3, 8 and 10 which are pointing out pair of nodes (11, 17), (13, 14) and (4, 13) are quite devious from this continuous pattern. We can examine that in pair of nodes (11, 17), the better topology is to designate feature 11 as class and feature 17 to be nominated as sink. This topology must possess a better scoring value but this is not correct for BDeu, MDL, Entropy and AIC. Moreover, for other two sets (13, 14) and (4, 13); none of the conventional veteran scoring function deliver correct prediction over the topology. However, throughout all of the experiments, the proposed measure I2S deliver a correct topology giving a perfect pattern such that a higher value of I2S means higher accuracy measures and lower false positive rate. In many cases, an increase in the fCLL value delivers a low value in TPR etc. Hence, there was no arguing in describing its score, although the shrewd readers may contact us for its completed detailed results of experimentation.

At this point, one can argue that the correct alignment of the values for prediction of the righteous (true) topology may be a stroke of fortune favouring the proposed measure I2S. Here the true topology indicates that topology which gives better discriminating result in a classification system. This motivates us to span this experiment over other benchmark dataset as shown by Figures 1 and 2. We in Figures 1 and 2 have shown a detailed summary of the results. The Figure 1 indicates the number of correct topologies revealed by various scoring function while the Figure 2 points out the number of in correct topologies drawn by well known and established scoring function along with our proposed measure.

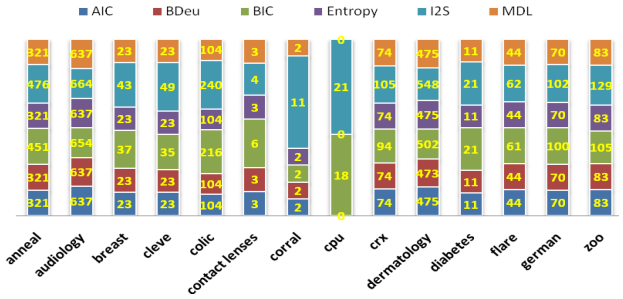

Figure 1. Comparison of scoring functions for correct (true) topologies of pairwise nodes.

The numeric data in both of the figures indicate the number of nodes with correct or incorrect topologies. For example, in case of dataset anneal, 321 number of pairs were drawn in such a topology which give maximum accuracy in case of AIC, Bdue, Entropy and MDL; whereas 451 and 476 number of pairs were drawn with correct topology in case of BIC and I2S respectively. The Figure 2 indicates the number of incorrect topologies drawn out using these scoring functions. For example, for the same dataset (anneal) the incorrect (wrong) topologies were minimum in case of I2S which is only 20. There were total 476 + 20 = 496 pairs in all selected such that all of the attributes can be covered in dataset anneal. It is evident that I2S deliver significantly better result in all of the fourteen datasets.

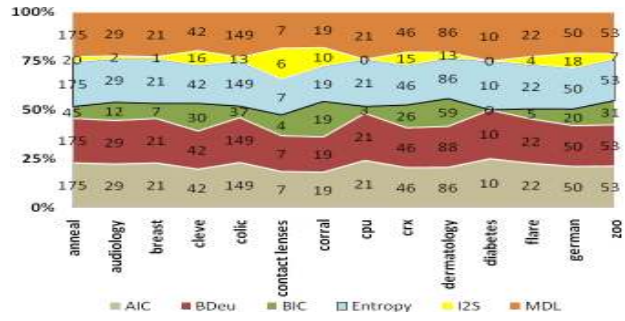

Figure 2. Comparison of scoring functions for incorrect (wrong) topologies of pairwise nodes.

We can draw two conclusions from a bird's eye view. Firstly the dataset 'CPU' completely gives highly favourable results for I2S and BIC where maximum number of incorrect topologies were reported for other scoring function. In this dataset, still I2S was leading scoring measure with highest percetage of true topologies. As far as the comparison to fCLL is concerned, we first obtained the source code available by the authors; the code is an extension of WEKA software. We made a slight change in the code

so that it can also, print out the final fCLL score at the exit of each cycle. However, we have restricted to the provision of its general result because detail of its scores come up with the conclusion that there was no particular fashion found in the scoring function of fCLL in comparison to accuracy measures. In many cases, an increase in the fCLL value delivers a low value in TPR. Hence, there was no arguing in describing its score. The potential reason behind this 'anomalous fashion' is that the said scoring functions evolves in the domain of TAN, C4.5, K-NN, SVM and LogR classifiers; but our experimentation uses K2 as searching algorithm.

At this point, it is useful to reiterate the lemma-4 which we mentioned previously. We described that "Given two nodes in BBN, I2S is prone to detect the best topology of two nodes". We can observe the same out of the experimentation that all conventional scoring functions exhibit minor or major deviations from the lemma 4. The scoring function fCLL is an exceptional notaion which in most of the cases gives no significant result. Hence, we can conclude that empirically a better scoring function is the one which is capable of exhibiting true topology of the given features. These experiments explain that I2S possess a relevant characteristic in development of topology of structure learning in Bayesian belief network. However, other mutual information measures can not indicate the correct topology of the dataset.

5. Conclusions

The most vital part in designing of a classifier is to bring forth discriminant functions within a feature space by means of utilizing a priori knowledge in a given training samples. Motivated from this fact, many classifiers were introduced by the experts of machine intelligence. It goes without say that in the approach of structure learning classification; Bayesian belief based classifier is par excellence with its prominent results. Moreover, application of mutual information in structure learning is not a novel idea as theoretically it was introduced some six decades ago. We pointed out a limitation of various scoring metric such as BDeu, AIC, Entropy, BIC, MDL and a recently introduced fCLL while introducing a new dependency measure in perspective of structure learning. This lay out the foundation of our hypothesis that our proposed measure is more suitable in employing a belief network. To validate our hypothesis, we first introduced the mathematical theory of the two measures indicating that if any measure is symmetric then it is not possible to truly define the orientation of topology of the nodes. This is true in case of Mutual information measure which is a symmetric measure in its nature. Later on, we picked six scoring metrics as mentioned above, and performed a series of exhaustive experimentation. The reason behind picking these six scoring measures is that five of them are implemented in WEKA and the sixth measure (fCLL) has been provided with online source code written as WEKA extension for delivering a reliable result in case of exercising belief network classifier. The empirical validation was carried out indicating that there is a difference in class imbalance measure for the results achieved on various dataset. The results clearly indicate that a higher value of I2S means higher accuracy in terms of various accuracy measures like precision, recall, f-measure and false positive rate. However, this conclusion is not uniform because the difference in the I2S for both of possible configuration is not in proportionate with corresponding difference in class imbalance metrics. There are two underlying possibilities: either a change in dataset or the second reason lies in the inability of proposed measures to explain it vigorously and in more depth. This left us with the option of improvement in the proposed measure so, that a proportionate difference in class imbalance metrics and I2S can be observed. We have left this work for future work. Another future direction for this work is to introduce a new breed of classifiers with quite reasonable performance exhibiting an improved accuracy. We also, plan to come up with a commercial classifier based on our introduced measure. Third future direction for this work lies in tailoring the measure to make it adaptable for a wide variety of dataset. Currently the measure is ideally suitable for multinomial (categorical) features only, whereas there is a need to encompass features other than multinomial dataset.

References

[1] Abbas A. and Liu J., "Designing an Intelligent Recommender System using Partial Credit Model and Bayesian Rough Set," *the International Arab Journal of Information Technology*, vol. 9, no. 2, pp. 179 - 187, 2012.

[2] Akaike H., "A New Look at the statistical Model Identification," *IEEE Transactions on Automatic Control*, vol. 19, no. 6, pp. 716 - 723, 1974.

[3] Aldrich J., "Correlations Genuine and Spurious in Pearson and Yule," *Statistical Science*, vol. 10 no. 4, pp. 364 - 376, 1995.

[4] Avilés-Arriaga H., Sucar-Succar L., Mendoza-Durán C., and Pineda-Cortés L., "A Comparison of Dynamic Naive Bayesian Classifiers and Hidden Markov Models for Gesture Recognition," *Journal of Applied Research and Technology*, vol. 9, no. 1, pp. 81 - 102, 2011.

[5] Blake C. and Merz C., "UCI Repository of Machine Learning Databases," available at: http://www.ics.uci.edu/~mlearn/,MLRepository.html, last visited 2012.

[6] Buntine W., "Theory Refinement on Bayesian Networks," *in Proceedings of UAI'*, Los Angeles, USA,pp. 52 - 60, 1991.

[7] Carvalho A., Roos T., Oliveira A., and Myllymäki P., "Discriminative Learning of Bayesian Networks via Factorized Conditional Log-Likelihood," *Journal of Machine Learning Research*, vol. 12, no. 7, pp. 2181 - 2210, 2011.

[8] Cheng J., Bell D., and Liu W., "An Algorithm for Bayesian Belief Network Construction from Data," *in Proceedings of AI & STAT'*, Lauderdale, Florida, pp. 83 - 90, 1997.

[9] Cooper G. and Herskovits E., "A Bayesian Method for the Induction of Probabilistic Networks from Data," *Machine Learning*, vol. 9, no. 4, pp. 309 - 347, 1992.

[10] Corder G. and Foreman D., *Nonparametric Statistics for Non-Statisticians: A Step-by-Step Approach*, John Wiley & Sons, New York, USA, 2009.

[11] Duwairi R., "Arabic Text Categorization," *the International Arab Journal of Information Technology*, vol. 4, no. 2, pp. 125 - 131, 2007.

[12] Gibbons J. and Subhabrata C., "Nonparametric Statistical Inference," *Marcel Dekker*, New York, USA, 2003.

[13] Grimmett G. and Stirzaker D., *Probability and Random Processes*, Oxford University Press, USA, 2001.

[14] Hall M., Frank E., Holmes G., Pfahringer B., Reutemann P., and Witten I., "The WEKA Data Mining Software: An Update," A*CM SIGKDD Explorations Newsletter*, vol. 11, no. 1, pp. 10 - 18, 2009.

[15] Heckerman D., Geiger D., and Chickering D., "Learning Bayesian Networks: the Combination of Knowledge and Statistical Data," *Machine Learning*, vol. 20, no. 3, pp. 197 - 243, 1994.

[16] Herskovits E., "Computer-Based Probabilistic Network Construction," *Doctoral Dissertation*, Stanford University, Stanford, CA, 1991.

[17] Jensen F. and Nielsen T., *Bayesian Networks and Decision Graphs*, Springer, New York, 2007.

[18] Lam W. and Bacchus F., "Learning Bayesian Belief Networks: An Approach Based on the MDL Principle," *Computational Intelligence*, vol. 10, no. 3, pp. 269 - 294, 1994.

[19] Meilaˇ M. and Jaakkola T., "Tractable Bayesian Learning of Tree Belief Networks," *Statistics and Computing*, vol. 16, no. 1, pp. 77 - 92, 2006.

[20] Messikh L. and Bedda M., "Binary Phoneme Classification Using Fixed and Adaptive Segment-Based Neural Network Approach," *the International Arab Journal of Information Technology*, vol. 8, no. 1, pp. 48 - 51, 2011.

[21] Schwarz G., "Estimating the Dimension of a Model," *Annals of Statistics*, vol. 6, no. 2, pp. 461 - 464, 1978.

[22] Srinivas S., Russell S., and Agogino A., "Automated Construction of Sparse Bayesian Networks from Unstructured Probabilistic Models and Domain Information, Uncertainty in Artificial Intelligence," *in Proceedings of the 5th Conference on Uncertainty in Artificial Intelligence*, Ontario, Canada, pp. 343 - 350, 1989.

[23] Suzuki J., "Learning Bayesian Belief Networks Based on the MDL Principle: An Efficient Algorithm Using the Branch and Bound Technique," *in Proceedings of the International Conference on Machine Learning*, Bally, Italy, pp. 356 - 367, 1996.

[24] Tlemsani R. and Benyettou A., "On Line Isolated Characters Recognition Using Dynamic Bayesian Networks," *the International Arab Journal of Information Technology*, vol. 8, no. 4, pp. 406 - 413, 2011.

[25] Velarde-Alvarado P., Vargas-Rosales C., Torres-Roman D., and Martinez-Herrera A., "An Architecture for Intrusion Detection Based on an Extension of the Method of Remaining Elements," *Journal of Applied Research and Technology*, vol. 8, no. 2, pp. 159 - 176, 2010.

[26] Wasserman L., *All of Nonparametric Statistics*, Springer, New York, USA, 2007.

[27] Wermuth N. and Lauritzen S., "Graphical and Recursive Models for Contingency Tables," *Biometrika*, vol. 70, no. 3, pp. 537 - 552, 1983.

Muhammad Naeem Research scholar at department of computer science, M. A. Jinnah University Islamabad Pakistan. His research area includes machine learning, text retrieval, graph mining, classification and data mining.

Sohail Asghar is Director/ Associate Professor at University Istitute of Information Technology at PMAS-Arid Agriculture University Rawalpindi Pakistan. He earns PhD in Computer Science from Monash University, Melbourne, Australia in 2006. Earlier he did his Bachelor of Computer Science (Hons) from University of Wales, United Kingdom in 1994. His research interest includes data mining, decision support system and machine learning.