# A REVIEW OF FEATURE SELECTION TECHNIQUES IN STRUCTURE LEARNING

Muhammad Naeem*, Sohail Asghar**

**ABSTRACT**

*In the last two decades, there has been significant advancement in heuristics for inducing Bayesian belief networks for the purpose of automatic distillation of knowledge from masses of data with target concepts. However, there are various circumstances where we are confronted to fix a set of most influencing variables in modelling of class variable. This arises in provision of confidence measures on set of variables used in the structure learning of data. In this study, we have tweaked empirical as well as theoretical aspects of various feature selection evaluators, their corresponding searching methods under six well known scoring functions in K2 which is a notable structure learning technique in Bayesian belief network. We have come up with some useful findings for overall computationally efficient approach among eleven evaluators. This analysis is useful in inducing better structure from the given dataset in imparting improved performance metric for Bayesian belief network.*

**KEYWORDS:** *Feature subset selection, Scoring Function, Bayesian classifiers, Benchmarking*

## 1. INTRODUCTION

Machine learning techniques are aimed towards automatic distillation of knowledge from machine readable information. However, their success is greatly influenced by the quality of the data under operation. Inadequate data with irrelevant as well as extraneous information restrict these techniques in narrow range of discovery with shortened precision. This phenomenon is termed as curse of dimensionality[1]. Feature Subset Selector (FSS) is a solution to the said problem. FSS can deliver reduced hypothesis space for searching with heightened performance. The primary objective of any FSS is to identify and eliminate superfluous information before the inception of learning phase[2]. Although there is already a survey paper found in the literature regarding the performance of various FSS; however our approach is quite different as explained in the forthcoming section.

Figure 1 is delineating the whole picture of the structure learning for which we shall describe each and every component of this framework shown in the Figure 1. The feature selection plays a very important role in achieving objective of the structure learning including wrong orientation, redundant features, extra edges and missing edges. In fact a careful selection of features can greatly improve the process of learning objection as shown in the Figure 1.

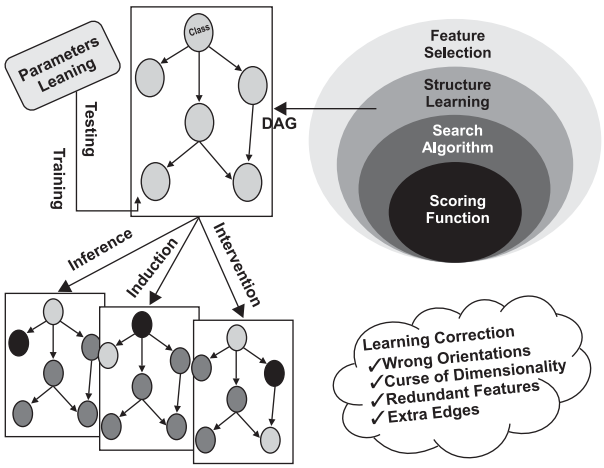


Figure 1: Structure Learning: A Broader Picture

The rest of the paper is organized into the following order. The next immediate section is our motivation for scripting this study. Section 3 is focused on the survey paper related to this study.In this section some quite relevant and useful literature review with a comparison to our analysis and review in this study is presented with detail information of Feature selection techniques. In Section 4, we have introduced Bayesian based classifier along with the evolution of core scoring function being used in BBN. In Section 5 feature selection evaluators and their types are discussed in detail. Section 6 brings the result of experimental methodology with discussion in detail.

* *Department of Computer Science, Mohammad Ali Jinnah University Islamabad Pakistan*
** *University Institute of Information Technology, PMAS-Arid Agriculture University, Rawalpindi Pakistan*

The section 7 explains the discussion on the results presented in three graphs followed by the last section of conclusion where we summarize some findings achieved in this study.

## 2. MOTIVATION

We particularly targeted only Bayesian Belief Network (BBN) classifier in detail. We examined each and every well established scoring function acting at the heart of the BBN. Nonetheless, we also include a recently introduced scoring function factorized conditional Log Likelihood (CLL)[3,4,5]. The previous study was related to comparison among three well know classifiers (Drugan et al., 2010), but our study explores single classifier with its array of central crux i.e. the scoring function. The previous study present result of seven evaluators but our study takes eleven evaluators[6]. It is a proven fact that BBN is a robust formalism and widely used technique. This motivates us to give a detailed impact of various FSS with respect to its conventional scoring function and any other acclaimed scoring function such as CLL. We found no specific literature review regarding this motivation; hence we come up the experimentation as given in this study. However, the comparison among different feature selection techniques using notable scoring function has not been addressed in the literature. This study is aimed towards provision of a user of BBN classification technique with providing an insight into given data by manifesting the relative merit of features of dataset.

## 3. LITERATURE REVIEW

To start with legacy literature review related to the topic in discussion, we shall discuss a review report presented one and half decade ago[7]. They presented work with a focus on categorization of available FSS techniques. They divided the evaluators into five groups: distance, uncertainty information, dependence, consistency, and classification error rate. They illustrated numerous dimensions for categorization / grouping. These grouping include ability to handle various data types, number of classes, small vs. large dataset, noisy vs. clean dataset and level of optimality. They performed these metrics on three synthetic dataset. However, the distribution of these synthetic dataset was biased. Therefore judgment for the correction of the evaluators was arguable. Nonetheless, such categorization was not a novel idea because feature selection techniques were also addressed in other dimensions before[7]. Doak et al.[8] categorizes the evaluators into data intrinsic measures, classification error rate and estimated or incremental error rate.

Another competitive and relevant review was introduced by Hall et al[9]. They exercised their experiment using weka software[10] in which they evaluated the comparison among seven evaluators using three common classifiers. Sayes et al[11], also produced a review on feature selection but restricted to only bioinformatics domain. They classified the techniques according to suitability, variety, usage and potential to sequence analysis and micorarray analysis. Although the pool of FSS techniques is becoming larger and larger[6,12,13]; nevertheless specific exhaustive review leading to a wealth of comparative report for Bayesian belief network's various scoring function is not addressed so far. We in this study have incremented useful information in these survey reports; moreover our analysis is more precise in tweaking BBN in particular.

## 4. BAYESIAN BELIEF NETWORK SCORING FUNCTION

It goes without say that great deal of research has been observed focusing on structure learning from data[14,15]. Bayes belief networks (BBN) have proved their robustness and efficiency in decision and reasoning under uncertainty for inference tasks. This effectiveness of BBN is grounded in terms of its capability for expressing structural and qualitative information about the domain of interest[16]. In BBN, structure learning has been addressed in two approaches; constrained based and scoring function inspired approaches. The later technique is more popular and intractable as compared to the first one[16]. The scoring function oriented approach which is essentially based on well established statistical principles, the whole structure is evaluated in terms of a score, the better the score, and the more reliable the network structure is. The score of the network in other words reflects how well the structure fits the underlying data; thus scoring function provides a pivot towards optimized structure learning.

Akaike Information Criterion (AIC) defined by Akaike[17] is first of its kind which was translated into

a scoring function as reported by Van et al[18]. Bayesian scoring estimation method[19] originally framed over network with hidden variables which otherwise culminates into well know BDeu score[20]. The other two notable scoring function include entropy based method[21] and Minimum Description Length (MDL) method[22]. Jensen et al., (2007) pointed out two essential properties for any BBN scoring function. The first property is the ability of any scoring function to balance the accuracy of a structure in context of structure complexity. The second property is its computational tractability. Recently Carvalho et al.[3,4,5] introduced factorized conditional log likelihood (CLL) and empirically proved it to be reasonable among other established scores. These scores formulates proposition for well motivated model selection criteria in structure learning techniques. However a noteworthy issue with employing these well established scores is that they are prone to intractable optimization problems. Chickering et al[23] argued that it is NP-hard to compute the optimal network for the Bayesian scores for all consistent scoring criteria. Another bottleneck with the performance metric of these scoring function is careful selection of features. It is already highlighted that feature selection can play an important role in evaluation of a classifier's performance. It is reported that little attention has been applied in evaluating the performance of BBN prior to its induction[13]. We can express our confidence that FSS is a key to estimation of performance of BBN prior to its induction phase in a real system.

## 5. FEATURE SELECTION EVALUATORS

The possible solutions to the curse of dimensionality can be trifurcated into three dimension. We shall discuss each one of them as below:

### 5.1. Feature Reduction

The first dimension is feature selection versus feature reduction. In feature reduction, new set of features are emanated from the existing set of features; in fact the actual features lose their identity at all. These techniques cater for sustaining maximum volume of information into a reduced number of newly born features. Latent Semantic Analysis and Principal Component Analysis both are data reduction techniques. In feature selection, only a sub set of the actual features is considered with the aim of rejecting the redundant and/or irrelevant to class features.

### 5.2. Feature Ranking

Feature ranking technically does not address the curse of dimensionality directly. However, there are some classifiers for which the initial feature ordering plays an important role in improvement of the classification accuracy. Naeem et al[24] has presented a useful insight into feature ranking along with introduction of a novel technique which is applicable for BBN and Random Forest classifiers. In general feature rankers are quite limited in their application. Firstly there are a few classifiers which are sensitive to feature ordering. It has been shown that there are situations when no feature or query variable can be surrendered but classification accuracy improvement is still imperative. Feature ranking or variable ordering becomes essential in such scenarios.

### 5.3. Feature Subset Selection

The third broad categorization is the set of techniques where an intelligent algorithm selects the most relevant features and shred all of the other query variables. Feature subset selection further comprises of three standard approaches; embedded approach, filter approach and wrapper approach. Although originally Kohavi et al,[25] introduced the binary category of filter and wrapper approaches; however, researchers argued that this category can be extended to third type known as embedded approach. The embedded approach is coined by the inherent nature of the underlying classification algorithm. The classification algorithm itself brings out the operation of feature selection under its criteria of supervised or unsupervised learning. OneR Attribute Evaluation is a notable example of such embedded approach where the logic of classification technique itself decides the selection of attribute at any specific level. In filter approach, features are selected a prior to the application of classification technique. Filter approach has nothing to do with the target classification technique in use. The filter approach rests on well defined statistically established principles such as pair-wise correlation, standard deviation etc. Majority of the FSS techniques belong to this category. In Table 1, except Wrapper Subset Evaluator, all of the techniques belong to this category. The wrapper approach is punched with the target classification technique which acts like a black box. Hall et al.[9], introduced another taxonomy marked by evaluation of individual or subset of features. It is

useful to present all of the available FSS technique in the table 1 under this category. We have presented only those evaluators which are available in weka[10]., 2009). This table will be helpful in the result section for analysis and comparison between both of the approaches. It is useful to give some precise insight into the general methodology of the evaluators which are in discussion in this study. We shall discuss each of them as following:

Gain Ratio Attribute Evaluator and Information Gain Attribute Ranking both are simple individual attribute ranking mechanism. In this technique, each attribute is assigned a score where the score is delineated by means of the difference of an attribute's entropy and its class conditional entropy. The difference between both of these entropies formulates the information gain for each of the attribute. Dumais et al[26] and Yang et al[27] reported that this uncomplicated technique is much suitable in case of text classification.

Relief Attribute Evaluator which is an individual attribute evaluation technique is more versatile as compared to its peer FSS because it can be operated on discrete as well as continuous data. Moreover, this technique is quite capable of handling noisy data. Originally it was introduced by Kira et al[28] for two classes only; however, it was improved for multiclass[29]. The central idea in this technique is identification of nearest neighbor from same as well as opposite class.

**Table 1: Taxonomy of Feature subset selection**

| Individual Attribute | Subset |
|---|---|
| Chi Squared Attribute Evaluator | Cfs Subset Evaluator |
| Filtered Attribute Evaluator | Classifier Subset Evaluator |
| Gain Ratio Attribute Evaluator | Consistency Subset Evaluator |
| Info Gain Attribute Evaluator | Cost Sensitive Attribute Evaluator |
| OneR Attribute Evaluator | Cost Sensitive Subset Evaluator |
| Relief Attribute Evaluator | Filtered Subset Evaluator |
| Symmetrical Uncertainty Attribute Evaluator | Wrapper Subset Evaluator |
| SVM Attribute Evaluator | |

CFS (Correlation-based Feature Selection)[30] is based on the evaluation of attributes subset; the success of this algorithm initiated a series of introduction of subset evaluators subsequently. The central crux of this technique relies on the idea of introducing such subsets which minimizes the inter-correlation and maximizing the intra-correlation. Here inter-correlation relates to the correlation among members of the subset and intra-correlation refers to the correlation to class variable. The rationale behind this technique is that the subset with attributes highly related to each other is prone to be poor predictor of the class.

Symmetrical Uncertainty Attribute Evaluator is restricted to discrete features only. This technique approximates the association score between discrete variables with respect to the class. Classifier Subset Evaluator and OneR Attribute Evaluator both are member of embedded class of FSS. The underlying logic behind OneR Evaluator is based on OneR classifier[31]. Chi Squared Attribute Evaluator is based on well established statistical measure for test of hypothesis where scoring value between each attribute and class is calculated for marking it as suitable or unsuitable feature for classification technique. Filtered Attribute Evaluator and Filtered Subset Evaluator both are filter based techniques. In both of these techniques, the attribute or set of attributes are evaluated by passing them through an arbitrary filter defined on the training dataset.

The general principle for Consistency-Based Subset Evaluation can be describes as the data is divided in such a way that the attributes with strong single majority class are separated from the other attributes[32,33]. This approach lay out the foundation for several FSS techniques.

Kohavi et al[25] introduced Wrapper Subset Evaluator. This breed of technique can never be operated independently. They always works keeping in view of the target data mining technique. This usually gives them an added advantage over their peer FSS techniques due to an enhanced interaction between the classifier's inductive bias and the searching mechanism. The estimated accuracy of the classifier is usu-

ally calculated by means of cross validation during the working of wrapper technique. The modified forward selection search is used to generate a ranked list of attributes. The only notable bottleneck of such techniques is increased computational cost specifically in case of large volume of attributes.

## 6. EXPERIMENTAL SETUP

We performed a series of exhaustive experiments using weka[10] which is a well know machine learning tool. The detail of the dataset used in the experiment is shown by the Table 2. The representative dataset for classification prediction problems followed by FSS were taken from machine learning database which is the data repository of university of California Irvine[34]. Majority of the dataset were having nominal discrete variables. The shrewd reader can notice that the dataset used in the study varies in cases, attributes and number of classes so that there should be no question of biasness for any specific scoring function in question. The detailed characteristics of these benchmark datasets is sum up in Table 2.

In order to give an unbiased comparison, it is compulsory to keep same parameters in the experimentation. The fixed parameters in the FSS evaluators are 'use full training set' in attribute selection mode. As we already mentioned that we tested eleven evaluator which include Symmetrical Uncertainty Attribute Evaluator (SU), Relief Attribute Evaluator (RL), OneR Attribute Evaluator (OR), Info Gain Attribute Evaluator (IG), Gain Ratio Attribute Evaluator (GR), Filtered Subset Evaluator (FS), Filtered Attribute Evaluator (FA), CfsSubset Evaluator (CF), Chi Squared Attribute Evaluator (CS), Consistency Subset Evaluator (CN), Wrapper Subset Evaluator (WP).

Search method for CfsSubset Evaluator is BestFirst while GreedyStepwise search method was used for Filtered Subset Evaluator and Consistency Subset Evaluator. RankSearch was employed for Wrap-

**Table 2: Dataset used in comparison of various feature subset evaluators.**

| Dataset | Cases | Attributes | Classes |
|---|---|---|---|
| Monk | 8416 | 22 | 7 |
| Chess | 3196 | 36 | 2 |
| Zoo | 101 | 16 | 7 |
| Dermatology | 358 | 33 | 5 |
| Mushrooms | 8124 | 22 | 7 |
| Soyabean | 266 | 35 | 15 |
| Nursery | 12960 | 8 | 5 |
| Flare | 1066 | 12 | 3 |
| Lymph | 148 | 18 | 8 |
| Vote | 435 | 16 | 2 |
| Anneal | 898 | 39 | 5 |
| Audiology | 226 | 70 | 24 |
| Autos | 205 | 26 | 6 |
| breast-cancer | 286 | 10 | 2 |
| Colic | 368 | 23 | 2 |
| credit-a | 690 | 16 | 2 |
| Diabetes | 768 | 9 | 2 |
| Glass | 214 | 10 | 6 |
| heart-c | 303 | 14 | 2 |
| Hepatitis | 155 | 20 | 2 |

| Dataset | Cases | Attributes | Classes |
|---|---|---|---|
| Hypothyroid | 3772 | 30 | 4 |
| Ionosphere | 351 | 35 | 2 |
| kr-vs-kp | 3196 | 37 | 2 |
| Labor | 57 | 17 | 2 |
| Relation | 20000 | 17 | 26 |
| primary-tumor | 339 | 18 | 21 |
| Segment | 2310 | 20 | 7 |
| Sick | 3772 | 30 | 2 |
| Sonar | 208 | 61 | 2 |
| Splice | 3190 | 62 | 3 |
| Vehicle | 846 | 19 | 4 |
| Vowel | 990 | 14 | 11 |
| waveform-5000 | 5000 | 41 | 3 |
| Australian | 690 | 15 | 2 |
| Cleve | 296 | 14 | 2 |
| Crx | 690 | 16 | 2 |
| German | 1000 | 21 | 2 |
| Satimage All | 6435 | 37 | 6 |
| Shuttle-Small All | 5800 | 10 | 6 |
| Pima | 768 | 9 | 2 |

per Subset Evaluator and Ranker search heuristic was used for all of the rest evaluators. The setup of experiment related to classification include selection14 of five conventional scoring function; Bayes, BDue, MDL, Entropy and AIC the theoretical detail, significance and evolution of these scores have already been discussed in previous sections. The sixth scoring function used in this experiment is CLL. The constant parameters for K2 are: 10 fold cross validation, maximum number of parents fixed to 5, initNaiveBayes and markovBlanketClassifier and randomOrder all were set to false. The status of randomOrder was important cause in all of the evaluators, the ranking of attribute is important and we know that K2 with different initial ordering always come up with different topologies; nonetheless a randomOrder setting of variables may lead to thwart the effect of FSS. Moreover, markovBlanketClassifier also refixes the structure after final stage of structure learning; such fixation of markovBlanket may yield a bias effect for actual retrieving of actual evaluation of FSS. While keeping in view of the same spirit, we also disable useADTree option and restrict the experiment to simpleEstimator with alpha value of 0.5 which is a default value for simpleEstimator of parameter learning. One noteworthy aspect related to WrapperSubsetEval using RankSearch heuristics is that this FSS evaluator always gives a ranking of attribute in a specific order and also a list of subset of features. We in this study take all of the features but keep them in the specific ranking order, such ordering as we already mentioned is very important if number of parents for any variable is kept more than one in drawing of DAG.

## 7. RESULT AND DISCUSSION

To measure the ability of different scoring function to be identified as the 'preferable choice', we adopted the simples measure "Accuracy" in the experimental result. Although there are other class imbalance measures of filters. However, we prefer to restrict to only "Accuracy" measure because firstly it was produced up to three or fourth decimal whereas the other measures were rounded off. This surely gives us a delicate difference between two values of accuracy.

The Figures 2, 3 and 4 all are representation of the comparison of evaluators on benchmark dataset. The figures illustrate how often each evaluator executes significantly better, worse, or non-effective at all. We shall discuss each of them one by one. The Figure 2 is showing the winning comparison. It indicates how many times an evaluator was successful in

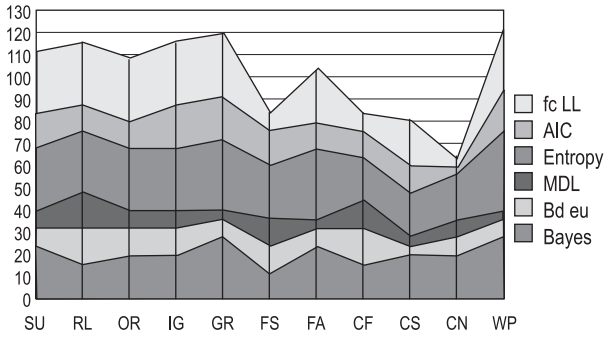


Figure 2: Winning comparison of evaluators with K2 scoring functions

achieving a better accuracy score under various scoring function. It is evident from the Figure 2 that Wrapper Subset Evaluator is an overall winner in the whole of the experiment. The Gain Ratio Attribute Evaluator enjoyed its status as runner up followed by Relief Attribute Evaluator and Info Gain Attribute Evaluator. The reason behind winning the Wrapper Subset Evaluator lies in the common assumption (monotonocity) stating that increasing the number of features usually increase the accuracy rate; although this is only a general assumption, we observed in numerous instances that a few of the attributes are required to be eliminated. However, if any evaluator did not pin point these features which are responsible for degradation in accuracy, the exemption of useful attributes drastically drops the accuracy factor of the classifier. Whereas when we analyze the other three evaluators, they are much worthy, cause these evaluators have put their best to come up with the best subset and apparently their performance is outnumbered by the other evaluators. We shall also examine another dimension of Figure 2 which is scoring function; among all of the seven scoring function, entropy scoring function occupied the largest share of the volume of the figure. This indicates that entropy based scoring function gives better result when used in all of the eleven evaluators. Another observation regarding entropy is that its performance was almost uniform under nine out of eleven evaluators where only a minor surge is observed in case of ChiSq.AttribEval and CfsSubsetEval. The scoring function CLL give better result in SymUncertAttribEval, ReliefAttribEval, OneRAttribEval, InfoGainAttribEval and GainRatioAttribEval. As we notice the scoring function which yields least; they are MDL and BDeu.

Figure 3 is an indication of statistical information which is related to *no less* and *no win*. In fact,

we observed that there are no many scenarios in which FSS neither perform well or bad. Two scoring functions MDL and BDeu are worthy enough to be mentioned in this category. Both of these are occupying significant volume of the graph. MDL keeps its

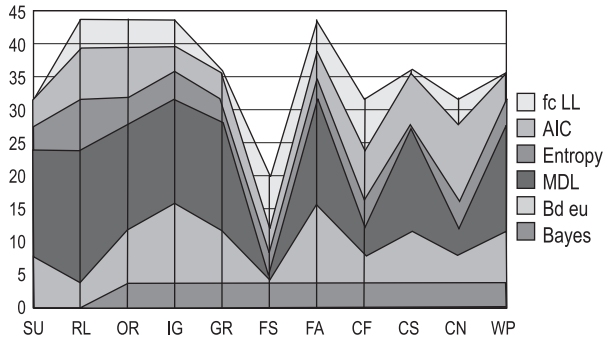


Figure 3: Comparison (no win, no loss) of evaluators with K2 scoring functions

behavior almost uniform except two evaluators FS and CF while BDeu also keep its behavior same except FS where it always goes for poor performance as shown by the Figure 4. When we look at the evaluator side, then we noticed four evaluators RL, OR, IG and FA which are higher overall in keeping their performance neutral. When we discuss the Figure 3 in perspective of Figure 2 then it can be concluded that RL, OR and IG have kept their status as either winner or neutral making them a good choice under any of the scoring functions.

Figure 4 depicts the loss rate of the evaluators. It indicates that FS, CF and CN in general did not deliver promising results and give many a times reduced accuracy. The same is true when we measure the performance of BDue and MDL. The figure 4 indicates that the least area is occupied by entropy whereas in Figure 2 the highest proportion is consumed by entropy. Although in Figure 3, its share is low, but based on the observation from three of the figures, we can conclude that entropy scoring function outnumbered when used in FSS evaluation. The runners up scoring function are cLL and Bayes scoring function. When we look at the worst scoring function, then BDeu performs poorly followed by MDL. The performance of AIC is quite intermediate in both of these extreme performances. On the other hand, if we conclude about the evaluators, then three of the Figures 2, 3 and 4 point out that WP, RL, IG and OR exhibited best. WP can be said as winner while the other three are almost equally runner up. The worst evaluator in the light of analysis achieved from three figures is conferred to CN and CS followed by CF.

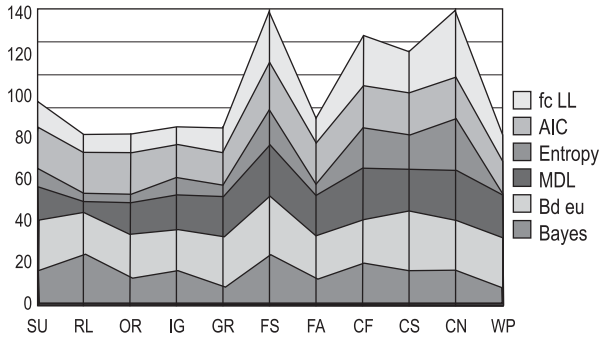


Figure 4: Lose comparison of evaluators with K2 scoring functions

## 8. CONCLUSION

When fabricating a BBN from dataset, the more or less superfluous variables included in a dataset; may bias the performance of a classifier. A high-dimensional dataset raises the likelihood that a classification algorithm may encounter to spurious patterns[1]. Earlier it was stated that the inclusion of an increasing number of query variables prone to increase the probability of inclusion of more information to distinguish between classes. However, this is technically incorrect as if the volume of the training dataset does not increase in proportionate with the inclusion of every new variable[1]. In this study, we analyzed the problem of curse of dimensionality in perspective of scoring functions which stays at the heart of any BBN. The objective of this study is to make machine learning / data mining community cognizant of the benefits, and in some situation even the requirement of utilizing feature selection methodologies in scope of Bayesian belief network. This study proposes a computational confidence on features selection methodologies of an induced model based on BBN structure learning. We can give some general recommendations regarding the selection of FSS evaluators where we termed WP, GR and IG as most suitable evaluators in the entropy scoring functions; although computational efficiency for WP has always been arguable. The empirical results also pointed out about the poor performance of CN and CF while BDeu scoring function did not perform well.

**ACKNOWLEDGEMENTS**

We express our thanks to anonymous reviewers whose cynical as well as appreciable review helped us a lot in preparing the final version of this study. Moreover, we are also grateful to Saira Gilani from department of computer science, Muhammad Ali Jinnah University Islamabad Pakistan for her valuable suggestion and assistance in form of proof reading.

**REFERENCES**

1. *Jensen, F.V. and T.D. Nielsen, 2007. Bayesian networks and decision graphs, Information Science and Statistics, Volume. ISBN 978-0-387-68281-5. Springer New York.*

2. *Hall M.A., and L.A. Smith, 1998. Practical Feature Sub set Selection for Machine Learning.*

3. *Carvalho, A.M., A.L. Oliveira and M.F. Sagot, 2007. Efficient learning of Bayesian network classifiers: an extension to the tan classifier. Proceedings of the 20th Australian joint conference on Advances in artificial intelligence, pp. 16-25.*

4. *Carvalho, A. M., T.T. Roos, A.L. Oliveira and P. Myllymäki, 2011. Discriminative learning of Bayesian networks via factorized conditional log-likelihood, Journal of machine learning research, Vol. 12, pp. 2181-2210.*

5. *Carvalho, A.M., 2009. Scoring function for learning bayesian networks, Technical report, INESC-ID Tec. Rep. 54.*

6. *Drugan M. M., and M.A. Wiering, 2010. Feature selection for Bayesian network classifiers using the MDL-FS score, International journal of approximate reasoning, 51(6), pp. 695-717.*

7. *Dash M., and H. Liu, 1997. Feature selection for classification, Intelligent data analysis, 1(1-4), pp. 131-156.*

8. *Doak, J., 1992. An evaluation of feature selection methods and their application to computer security", Technical report, (1992), Davis, CA: University of California, Department of Computer Science.*

9. *Hall M.A., and G. Holmes, 2003. Benchmarking attribute selection techniques for discrete class data mining, IEEE Transactions on Knowledge and Data Engineering, 15(6), pp. 1437-1447.*

10. *Hall M.A., E. Frank, G. Holmes, B. Pfahringer, P. Reu temann, I. H. Witten, 2009. The WEKA Data Mining Software: An Update, ACM SIGKDD Explorations, Volume 11, Issue 1.*

11. *Saeys Y., I. Inza, P. Larrañaga, 2007. A review of feature selection techniques," BIOINFORMATICS, Vol. 23 no. 19 pp. 2507–2517.*

12. *Samb M.L., F. Camara, S. Ndiaye, Y. Slimani, and M.A. Esseghir, 2012. A Novel RFE-SVM-based Feature Selection Approach for Classification, International Journal of Advanced Science and Technology, Vol. 43, June, pp. 27-36.*

13. *Khor K.C., C.Y. Ting, and S.P. Amnuaisuk, 2009. From feature selection to building of Bayesian classifiers: A network intrusion detection perspective, American Journal of applied sciences, 6(11), pp. 1949-1960.*

14. *Buntine W., 1996. A guide to the literature on learning probabilistic networks from data, IEEE Transactions on Knowledge and Data Engineering, 8(2), pp. 195-210.*

15. *Heckerman D., 2008. A tutorial on learning with Bayesian networks, Innovations in Bayesian Networks, pp. 33-82.*

16. *Guo Y., and D. Schuurmans, 2012. Convex structure learning for Bayesian networks: Polynomial feature selection and approximate ordering, arXiv preprint arXiv: 1206.6832.*

17. *Akaike, H., 1974. A new look at the statistical model identification.* Automatic Control, IEEE Transactions on, 19*(6), 716-723.*

18. *Van A.T., and R. Greiner, 2000. Model selection criteria for learning belief nets: An empirical comparison, Machine learning-international workshop then conference, pp. 1047-1054.*

19. Cooper G.F., and E. Herskovits, 1992. A Bayesian method for the induction of probabilistic networks from data, Machine learning, 9(4), pp. 309-347.

20. Heckerman D., D. Geiger and D.M. Chickering, 1995. Learning Bayesian networks: The combination of knowledge and statistical data, Machine learning, 20(3), pp. 197-243.

21. Herskovits E.H., 1991. Computer-based probabilistic network construction, Doctoral dissertation, Medical information sciences, Stanford University, Stanford, CA.

22. Suzuki J., 1999. Learning Bayesian belief networks based on the MDL principle: An efficient algorithm using the branch and bound technique, IEICE Transactions on Information and Systems, 82(2), pp. 356-367.

23. Chickering, D. M., D. Heckerman, and C. Meek, 2004. Large-sample learning of Bayesian networks is NP-hard, The Journal of Machine Learning Research, 5, pp. 1287-1330.

24. Naeem, M., and S. Asghar, 2013. A Novel Feature Selection Technique For Feature Order Sensitive Classifiers, Anale. Seria Informatica. Annals. Computer Science Series, Vol. XI fasc. 1, Pp. 31-38.

25. Kohavi, R. and G.H. John, 1997. Wrappers for feature subset selection, Artificial intelligence, 97(1), pp. 273-324.

26. Dumais, S., J. Platt, D. Heckerman, and M. Sahami, 1998. Inductive learning algorithms and representations for text categorization. In Proceedings of the seventh international conference on Information and knowledge management, pp. 148-155, ACM.

27. Yang Y., and J.O. Pedersen, 1997. A comparative study on feature selection in text categorization, Machine learning-international workshop then conference, pp. 412-420, morgan kaufmann publishers, inc.

28. Kira K., and L. Rendell, 1992. A Practical Approach to Feature Selection, Proc. Ninth Int'l Conf. Machine Learning, pp. 249-256.

29. Kononenko, I. 1994. Estimating attributes: analysis and extensions of RELIEF, Machine Learning: ECML-94 pp. 171-182. Springer Berlin/Heidelberg.

30. Hall M.A., 2000. Correlation-Based Feature Selection for Discrete and Numeric Class Machine Learning, Proc. 17th Int'l Conf. Machine Learning (ICML2000).

31. Holte R.C., 1993. Very simple classification rules perform well on most commonly used datasets, Machine Learning, Issue (11) pp. 63-91.

32. Almuallim H., and T.G. Dietterich, 1991. Learning with many irrelevant features, Proceedings of the ninth National conference on Artificial intelligence, Vol. 2, pp. 547-552.

33. Liu H., and R. Setiono, 1996. A probabilistic approach to feature selection-a filter solution, machine learning-international workshop conference, Morgan kaufmann publishers, inc. July, pp. 319-327.

34. Blake, C., E. Keogh and C.J. Merz, 1998. UCI Repository of Machine Learning Data Bases, Univ. of California, Dept. of Information and Computer Science, Irvine, CA