A NOVEL FEATURE SELECTION TECHNIQUE FOR FEATURE ORDER SENSITIVE CLASSIFIERS

Muhammad Naeem

Mohammad Ali Jinnah University Islamabad Pakistan, Department of Computer Science

Sohail Asghar

PMAS-Arid Agriculture University, Rawalpindi Pakistan, UIIT

ABSTRACT: In this study, we introduced a novel feature ranking technique applicable to two well known classifiers Bayesian Belief network and Random Forest as both of these classification systems have been shown to be sensitive to the initial ordering of the features. We have illustrated that improvement in classification can be obtained even without ceding variables for feature (attribute) ranking sensitive classifiers. We also performed a comparison between Bayesian Belief network and Random Forest classification approaches in the well known feature subset selection and feature ranking problem. The proposed technique Polarization Measure (herein known as PM) is originated from within joint probability to discover the degree of explanation made by first feature (attribute)'s state to explain the other feature's state. The technique has significantly better well performed in Bayesian belief network and better in random forest classifier in comparison to five feature ranking techniques and three well established feature subset selection techniques.

KEYWORDS: random forests algorithm; machine learning; Bayes structure learning; ranked features.

1. INTRODUCTION

In the domain of data mining, classifier's ability to predict with high accuracy is vital. The performance of classification system has been reported sensitive to the underlying characteristics of the data. It is reported that the performance of a classifier system is a function of discriminative variables. Numerous feature subset selection systems have been reported in last two decades; however no universal technique has been introduced to cater each and every kind of data which is applicable to every classification system. Through whole of this study, we have used the terms variable and feature as interchangeable to each other.

It is a preliminary requirement for any classification system to get its input 'prepared'; here the 'prepared' denotes that the input must be presented in the form of binary, nominal or categorical feature values. Although, feature selection is found useful for every classifier and this leads to the emergence of numerous taxonomies in literature; but there are situations when the user of classifier does not want to surrender any features but improve the accuracy of the classifiers. Moreover, it is argue-able that some classifiers are quite sensitive to the order of the variables supplied as basic input to the classifier. In such scenarios, feature order sensitive classifiers become more important. These include Random Forest, Naïve Bayes Belief network, PART, TAN etc. We in this study have analyzed that two classification systems have been found significantly sensitive to the order of variables / features involved. These include Naïve Bayes Belief Network and Random Forest. Feature Ranking (FR) algorithms like feature subset selection algorithm establishes the relevance of an attribute compared to the class, however, there is no question of dropping any feature.

When we talk in the perspective of the feature ordering rather feature selection under the gist of improvement in the classification then we can divide feature selection techniques into two categories. Ranking based feature selection techniques and Subset based feature selection techniques. Some notable techniques which can work be grouped into first category include Gain Ratio, Info Gain, Relief Attribute, Symmetrical Uncertainty Attribute, Chi Squared Attribute. While some well known techniques in the second category include Correlation based Feature Subset selection using, Consistency Subset and Filtered Subset. We have made comparison of our proposed technique to all of these techniques. Although we are not the first to define a concept of feature ranking in the context of classification system yet there are numerous contribution made by us in this study. We can briefly summarize the originality of the this study revolving around following contributions

- We have introduced a novel measure of coherence between two features (herein called as Polarization Measure (PM)).
- We have developed a new heuristics which used the PM at its heart while giving an optimized ranking list of the variables of dataset. Here

optimized stands for the best ranking of the features for the classifier

- The introduced technique is not only efficient but also giving better result as compared to its peer techniques.
- The introduced technique is quite scale able as well as stable to very large and very small dataset.
- We have shown empirical results for the comparison between well known ranking feature selection techniques with some recommendations.

The rest of the paper is organized into five sections. Section 2 and 3 covers some basic background of random forest and BBN classifier in perspective of their importance, core functionality and sensitivity towards feature ranking. In section 4, we have introduced our proposed technique. In section 5, we have discussed the empirical results supporting our claims made in this section. In section 6, we have delivered some useful conclusion after discussion on the topic in this study.

2. RANDOM FOREST CLASSIFIER

Random Forest classifier which is a homogeneous ensemble was initially introduced by Breiman [Bre01]. It takes un-pruned decision trees such that each node of the tree with best feature from a randomly selected subset out of all features is chosen. The data sampling used in this process is bootstrap in which sampling is performed with replacement from the original dataset. The un-pruned trees are built for reducing bias while the randomization is meant for maintaining high diversity between trees in the forest. We shall recall that here Bias denotes the systematic error term which is independent of the underlying learning sample; whereas the variance is the error caused due to the variability of the model in the learning sample randomness. The crux of this classifier revolves around it voting. Decisions are determined through simple voting. This approach has been recognized a well known, well established successful ensemble methods. The generalization error of a forest depends on the strength of the individual trees in the forest and upon the dependence between them. Random forests approach delivers high classification accuracy as compared to numerous well established classification techniques such as AdaBoost [FS95] and SVM [Vap99]. The reason behind its better performance roots in its abililty of being robust to noise, void of over fitting problem and its improved time complexity [Die00]. One notable characteristic is its high efficiency over significantly large dataset. The application of RF classifier has been reported in diversified domain of interest. Some include identifying curvilinear structure of mammograms in the domain of biomedical image processing [B+11], Genomic selection [OPS11], machine fault diagnosis [YDH08], Natural Language Processing [KP08] and many more.

Breiman [Bre01] indicated that Random Forest is akin to generate error rates almost at the same level as in case of Bayes rate in various domain of interest. However, [Rob04] pointed out that improvement in accuracy in some domain is possible either through application of a combination of various feature selection criteria to decrease correlation in the forests, or in other way, substitution of majority voting by means of locally weighted voting. This motivates us that there is a significant margin to apply any of the technique to improve the accuracy of Random Forest classifier. Ozcift [Ozc12] introduced a wrapper feature subset evaluator which uses Random Forest as its kernel. They exercised the evaluator over four dataset and presented improved results in comparison to fifteen classifiers. However, their technique surrender very large fraction of actual dataset. Such technique may become argue able in situations where the users have the intention to utilize all or large proportion of the original features. Menze et al. [M+11] introduced a version of Random Forest namely oblique Random Forest (oRF). It was shown that oRF is built out of multivariate trees which precisely learn optimal split directions at internal nodes employing linear discriminative models instead of applying the random coefficients in RF. Moreover it was also observed that it is optimized in classification ROC Area for those dataset which have tighter correlation among features; however its overall usefulness is limited to only binary classes. Breiman [Bre01] pointed out that the un-pruned trees in random forest are drawn for reducing bias. Earlier to it, Buntine [Bun92] presented Bayesian based classification algorithms for the purpose of tree averaging to shorten the variance in learning procedures. Later on, numerous techniques used Naïve Bayes theorem in their discriminant functions.

3. BAYESIAN NETWORK CLASSIFIER

A Bayesian network encodes the probabilistic relationship with the given data to model in perspective of a class variable. This model is capable of not only describing the data but can also produce new instance of the features possessing same statistical properties as in the original dataset. A BBN encodes the probability distribution of variables know as P(x) such that a vector of variables $X = \{X_1, X_2, X_3, \ldots. X_d\}$. We can write it as tuple of (M, Θ) where M indicates model comprised of Directed Acyclic Graph (DAG) and Θ indicates the maximum likelihood parameter for the model M. The objective

function in structure learning is to obtain the optimized model among a super exponential number of possible models. In BBN, the factorization of the probability distribution of variables is carried out by a Bayesian network induced by a cheap local search mechanism like K2, TAN etc. K2 algorithm [CH92] draws a Belief network out of discrete features with the underlying conditional independence. One notable assumption which is quite sensitive to performance of this algorithm is ordered set of attributes. This assumption of preordering attributes asserts that for any attribute, no attribute can be its parent provided if it comes later in order after that specific attribute. It means the last attribute in the ordered list may be tested against all of the other attributes for being its possible parent; while the first attribute must not have any parent at all. This first attribute will be treated a root node in DAG. The algorithm employs a greedy heuristic in which parent set for each node is developed with the underlying criteria whether addition of a node in parent list maximize the probability structure of the whole network or not. The crux of this algorithm revolves around this mathematical formulation as below:

$$f(i,\pi_i) = \prod_{j=1}^{qi} \frac{(r_i - 1)!}{N_{ij} + r_i - 1)!} \prod_{k=1}^{ri} \alpha ijk\,!$$

We shall refer the research by Cooper et al., [CH92] for its complete explanation and in depth working functionality. But here in this study we are concerned only with highlighting this fact that ordering of attributes plays an important role in K2 algorithm. The effective application of K2 algorithm has been shown in many research studies [CSG04, Hsu04, Car09]. The problem of attribute ordering in BBN can be tackled in two ways, either the knowledge of a domain expert may determine the preferred ordering list or otherwise we are left with automatic technique in which some machine intelligent algorithm defines the appropriate ordering such that quality of the learnt structure be higher than other possible structures out of a very large space of structures.

Some related literature review includes [Z+04]; Zio et al. [Z+04] argued the ordering be made according to reliability. The term reliability was a function of low quantity in missing values with considerable higher level of availability of data. It is argue-able that such an approach is limited to only dataset possessing missing records. Friedman et al., [F+00] shows that search space for asserting most suitable ordered list of features is less in size in comparison to search for optimized Bayesian structure ; where the later was proved to be NP hard [CMH03]. A very much related technique was introduced by Estevam et al., [E+07]. They showed that chi squared statistical test [DeG70] and information gain [Qui93] metric are quite useful to evaluate a better variable ordering scheme using the K2 heuristics for learning BBN out of data. They illustrated their empirical results over a few dataset. In this study, we introduced an approach based on a novel measure for feature ranking suitable for BBN Classifier learning and random forest classifiers context and while doing so, we have also made comparison to chi square and information gain based techniques.

4. TOWARDS A NOVEL FEATURE RANKING MEASURE

The central issue in supervised data mining techniques is associated with induction of discriminant model identifying a given instance of an object mapping into a specific class. The induction of the classifier requires that each object is to be enumerated by means of an array of variables. With the advent of advanced computational technologies, data with numerous features is quite a norm. In this scenario, a fundamental axiom is built: the degree of usefulness of all of the available variables for inducing the optimized model. Here the data mining community usually comes up with the idea of selecting the best subset; however there might be the situation when not even a single feature can be surrendered while still keeping the target of optimized induction model.

It is useful if we develop mathematical expression by characterizing feature ranking problem in the context of machine learning. We begins with T= D(F,C) as a sampling space or also known as training dataset in which there are g features and h instances; the set of features can be expressed as F = $\{f_1, ..., f_g\}$ and the dataset instances can be represented as D=$\{d_1, ..., d_h\}$. Moreover C = $\{c_1, ..., c_n\}$ refers to the set of tagged labels so here in this case they are classes. For each instance $d_x \in D$, it can be denoted as a vector of features, i.e., $o_x=(v_{x1}, ..., v_{xg})$, where v_{xk} is the value of o_x related to the feature f_l. Given a training dataset T= D (F, C), the task of learning algorithms for classification is to induce a hypothesis $h_0 : F_l \rightarrow C$ from T, where Fl is the value domain of $f_l \in$F.

After this brief introduction we shall move towards inscribing polarity measure between a feature and class attribute. Let the distinct state of the feature are expressed as F_1 = {1, …. m} while the distinct states of the class attribute are C = {1,…. n}. The value of h is already defined as the count of instances in the dataset. Now we denote a_{ij} as the joint probability among feature F_l and Class C, then Polarity Measure (PM) can be mathematically denoted as below:

$$PM\ (f_l, C) = \sum_{i=1}^{m} \left\lfloor \left(\max_{i=1}^{m} \arg \left[a_{ij} \right]_{j=1}^{n} \right) \Big/ h \right\rceil \qquad (1)$$

The above equation is in fact a mathematical expression to find the polarity measure between any of two attributes; the only difference is that the class is placed at the second position. For the purpose of comparison it is compulsory to calibrate the value of PM between 0 and 1. The equation above gives a value from 0 to 1 in the following particular situations.

$$PM_{\max} \leftarrow 0 \therefore \left\lceil \left[\overset{\Uparrow}{a}_{ij} \right]_{i=1,j=1}^{m,n} - \left[\overset{\Downarrow}{a}_{ij} \right]_{i=1,j=1}^{m,n} \right\rceil \rightarrow 0 \qquad (2)$$

$$PM_{\max} \leftarrow 1 \therefore \left\lceil \left[\overset{\Uparrow}{a}_{ij} \right]_{i=1,j=1}^{m,n} - \left[\overset{\Downarrow}{a}_{ij} \right]_{i=1,j=1}^{m,n} \right\rceil \rightarrow h \qquad (3)$$

a_{ij} : Joint probability with feature 1 in i[th] state and feature 2 in j[th] state.

$\overset{\Downarrow}{a}_{ij}$: Minimum joint probability among all of the possible space

$\overset{\Uparrow}{a}_{ij}$: Maximum joint probability among all of the possible space

In the next two step, we shall calculate the difference of the PM value such that the first value is PM (f_i, C_l) from feature to class where as the other value of PM(C_l, f_i) is obtained after swapping the position of the class node and the feature node. The net value is also divided by the later so as to find the net discirminant effect. All of these discriminant values are sorted in asending order. This will give us a list of feature which is ranked. We demonstrated that the polarization measure can explain the state of the class by means of expressing joint probability in a specific manner.

4.1 Asymptotic Analysis

The equation for the PM given above indicates that it needs a single scan of all of the database transactions. It means the time complexity for this measure is O(n). However inside three iterations, we need to update the information into hash table such as:

```
for h = 1 ..... sizeof (D)
    for i = 1..... m
        for j = 1..... n
            hashtable a_ij ← D(v_i, c_j)
        end for
    end for
end for
```

We know that Hash table time complexity is measured in terms of amortized analysis [AN07]. The amortized analysis is different from average-case performance analysis because of using probability in its analysis [AN07]. The best time complexity of open hashing in which a single array element can store any number of elements is O(1). Whereas the worst time complexity is O(h) although we can improve it to O(Log(h)) by using a balanced binary tree for each bucket. Hence, we can conclude that the total time complexity for the PM = O(h) x O(log(h)) = O(hxLog(h)) in worst case while it is Θ(h) X Θ(1)= Θ(h) in best case.

5. EMPIRICAL VALIDATION

A number of benchmark datasets have been used for the evaluation in this study. These include dataset with binary classification problems as well as multivariate classification problems obtained from the UCI data repository [TC09]. Table 1 is indicating an overview of these dataset in which attributes count, number of rows (cases) and classes are shown. It is preferred if we select dataset with variety of information under these categories to avoid any bias results in favor of a specific technique. All of these dataset were discretized using weka [H+09].

To measure the ability of various feature ranking techniques along with their respective searching algorithm can be termed as the 'preferable choice'; we adopted the simple standard classification measure "Accuracy" in this experiment. Although other robust class imbalance measures are also available; however, we prefer to restrict the experimental evaluation to the "Accuracy" measure because firstly in weka system, it is produced up to three or fourth decimal level whereas other measures are rounded off. This level of precision certainly delivers a delicate difference between two values of accuracy. Secondly, there is an inherent monotonic relationship between all of these classification imbalance characteristics such that if accuracy of a classification result is higher for a specific dataset under certain pre requisites then all of the other class imbalance characteristics like true positive rate, recall, precision etc would also be higher in the same fashion. We may conclude that for sake of comparison; any of the class imbalance measure can

be adopted and we in this experiment pick the simplest standard measure known as 'Accuracy' measure.

The setup of experiment related to classification includes selection of searching algorithm. The search algorithm during the structure learning is fixed to K2 wherein the parameters for K2 were default parameters in which initialize as Naïve Bayes is an important parameter. If it is set to false then it implies an empty network is to be used as initial network structure but if it is set to true then it means the initial network being used for structure learning will be a Naive Bayes Network, in which there will be an arrow from the node to each other node.

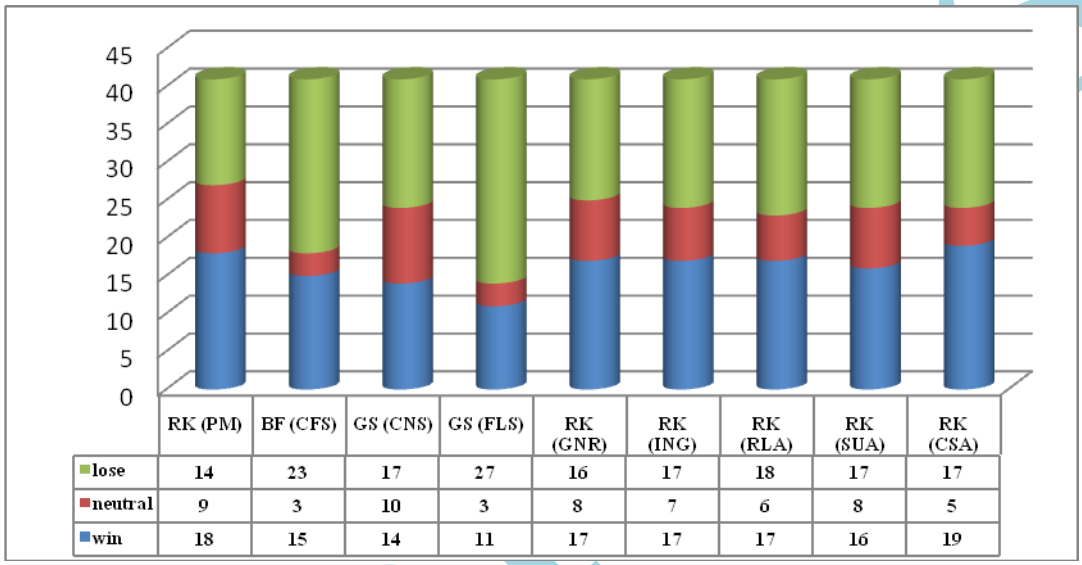
Table 1. Dataset used in comparison of various feature subset evaluators

Dataset	Attrib	Cases	Classes	Dataset	Attrib	Cases	Classes
Anneal	39	898	5	letter	17	20000	26
Audiology	70	226	24	lymphography	19	148	8
Australian	15	690	2	mammographic	6	830	2
balance-scale	5	625	3	mofn	11	1324	2
Bupa	7	345	2	monk-2	7	432	2
Car	7	1728	4	mutagenesis-atoms	12	1618	2
#chess	37	3196	2	nursery	9	12960	5
Colic	23	368	2	pima	9	768	2
contact_lenses	5	24	3	#primary-tumor	18	339	21
#crx	16	690	2	satimage	37	6435	6
Diabetes	9	768	2	segment	20	2310	7
#eastWest	26	213	2	shuttle	10	5800	6
Flare	13	1066	3	sick	30	3772	2
Glass	10	214	6	soybean-large	36	266	15
glass2	10	163	2	tae	6	151	3
Haberman	4	306	2	tic-tac-toe	10	958	2
hayes-roth	5	160	3	titanic	4	2201	2
#hepatitis	20	155	2	vehicle	19	846	4
kr-vs-kp	37	3196	2	vote	17	435	2
#labor	17	57	2	waveform	22	5000	3
led7digit	8	500	10	zoo	17	101	7

We set the default value which is true. The second setting is related to implementation of markove blanket classifier. This setting indicates that when complete structure learning is achieved then it is tested whether every node is a part of markove blanket for the nominated class and if any node is out of this setting then a correction is made. The default setting was set to false in this case. Moreover, if it is set to true then a true effect of the feature ranking cannot be observed; this left with the option of proceeding with default setting of 'false' in the underlying experiment. The other setting is maximum number of parents. In fact this setting is directly related to the computational efficiency of the structure learning in BBN. It is already an established fact that increasing this value may lead to exponential rise in time complexity of the algorithm. We chose a value such that we can get result for all the dataset with small or large number of attributes; the value we chose in our experiment is four. Another setting in K2 was related to random ordering of the input feature which was also set to false. The status of random Order is nontrivial because in all of the evaluators, the ranking of attribute was about to be tested and if this value is set to a function of randomness then effect of the feature ranking and feature selection algorithm cannot be judged correctly. In fact, random ordering of the features and markove Blanket Classifier values has no effect if number of parents is restricted to only one. The default scoring function used in BBN is Bayes Information Criterion and the same was considered in the experiment.

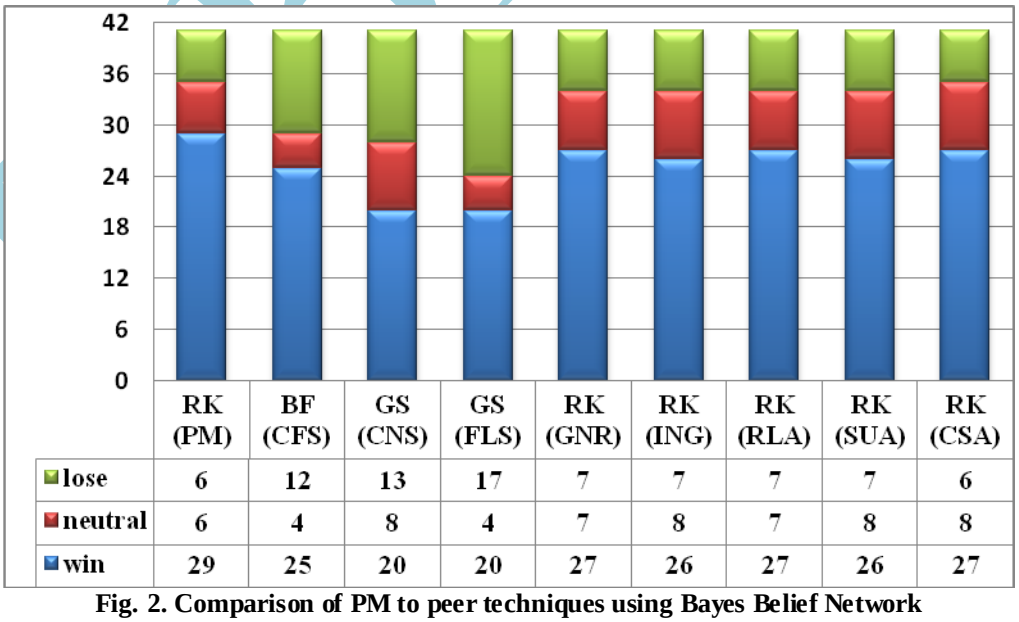
The theoretical detail, significance and evolution of Bayes Information Criterions as a scoring function has already been enumerated in previous sections. While keeping in view the same trend of default setting, we also disable use AD Tree option and restrict the experiment to simple Estimator with alpha value of 0.5 which is a default value for simple Estimator of parameter learning. Other fixed parameters for BBN classification include 10 fold

cross validation. Some standard setting used in this experiment related to RF include the maximum depth of the trees set to unlimited, the number of features to be used in random selection was set to use all features as we are interested in ranking of features, the number of trees to be generated was set to 10 which is a default setting. The random number seed to be used is set to its default value of 1. The rationale behind using these default values is that the technique suggests these optimized parameters for best results so there was no argue in altering these default values.
The design and comparison strategy used in the experiment is such that: we chose five feature ranking algorithm (evaluators). These include Gain Ratio evaluator (GNR), Info Gain evaluator (IGN), Relief Attribute Evaluator (RLA), Symmetrical Uncertainty Attribute evaluator (SUA) and Chi Squared Attribute evaluator (CSA). All of these evaluators used ranker as their searching heuristics. Apart from these feature ranking evaluators, we also used three well known feature selection techniques being used with their optimized searching algorithm. These include Correlation based Features Selection (CFS) evaluators using Best First (BF) searching algorithm [Hal00], Consistency Subset (CNS) evaluator and Filtered Subset evaluator (FLS) both using Greedy Stepwise (GS) searching algorithm.



	RK (PM)	BF (CFS)	GS (CNS)	GS (FLS)	RK (GNR)	RK (ING)	RK (RLA)	RK (SUA)	RK (CSA)
lose	14	23	17	27	16	17	18	17	17
neutral	9	3	10	3	8	7	6	8	5
win	18	15	14	11	17	17	17	16	19

Fig. 1. Comparison of PM to peer techniques using Random Forest classifier



	RK (PM)	BF (CFS)	GS (CNS)	GS (FLS)	RK (GNR)	RK (ING)	RK (RLA)	RK (SUA)	RK (CSA)
lose	6	12	13	17	7	7	7	7	6
neutral	6	4	8	4	7	8	7	8	8
win	29	25	20	20	27	26	27	26	27

Fig. 2. Comparison of PM to peer techniques using Bayes Belief Network

Empirically, it is seen from figure 1 and 2 that PM feature ranking strategy improves RF classification performance in eighteen dataset while it does not affect the performance in 9 cases albeit it give poor results in fourteen datasets. This result is comparatively better in comparison to its peer techniques. Moreover, it also indicates that there are situations when choosing a part of whole space of features may lead to poor performance of the classifier as such in the case of CFS using its

optimized BF searching algorithm. More or less, we can also argue in the same way about Classifier subset which acts in the capacity of wrapper techniques in which it tests the intended classifier during the selection of most influential features. In particular our proposed technique produced significantly better performance on BBN classifiers to majority of its peer techniques and better but not significant improvement performance on RF classifiers for six out of eight feature selection techniques.

The figure 2 is indication of amelioration or deterioration of BBN classification accuracy. The figure 2 also supports our proposed technique in general where in 28 dataset the improvement was observed while in 6 dataset, the accuracy was reduced and in 6 dataset the accuracy was neither improved nor reduced. These dataset have been marked by # sign in table. A careful examination of table-1 reveals that the empirical reason for reduction in accuracy for PM in 6 dataset seems to be the small size of dataset albeit we cannot generalize it. The figure 1 and 2 also points out that BBN classifiers is more responsive to change in feature ordering scheme whereas RF classifier has not shown as significant response as compared to BBN.

6. CONCLUSION & DISCUSSION

Classification is an important technique in expert systems to support the domain experts to identify knowledge out of large volume of data. The performance of such expert systems is greatly influenced by the accuracy of the core classifier used in the underlying design of the system. In pursuit of improvement in the accuracy of classification system, one significant and initiating step is to build the efficient feature reduction strategy. However, there might be situations where the domain expert is interested in retaining all of the features. At this point, idea of feature ranking becomes more useful and interesting. The concept of feature ranking is limited to those classifiers which are quite sensitive to the initial ordering of the input features. Although some well known feature ranking techniques are already available; however, we have shown that improvement in feature ranking is possible while addressing the enhanced accuracy of the classifier. In feature ranking, our proposed technique based on introduced measure Polarization Measure proves itself efficient in Random Forest and notable Bayesian Belief Network classifiers with the experimental results presented. Previously, chi square and information gain feature ranking algorithm were shown to be effective in producing better results. However previous results [E+07] were having asymptotic complexity of $O(n^2)$, but we have improved it up to $\Theta(nxLog(n))$ in worst case. We in this study have shown our result to comparatively better to not only chi square and info gain but also some other well known ranking techniques. Moreover, previously the technique was restricted to only BBN classifier but we applied our technique to both BBN and RF. Thus, achieving better results in circumstances with inclusion of all of the features by a machine learning technique seems to justify the proposed technique. In future work, we can extend this work in such a way that an ensemble would be introduced which will comprise of all of these six ranking features including the proposed technique based on proposed measure 'Polarization Measure'.

7. REFERENCES

[AN07] **A. Asuncion, D. Newman** - *UCI repository of machine learning databases, university of California*, deptartment of information and computer science, Irvine, CA, 2007: http://www.ics.uci.edu/~mlearn/{MLR}epository.html.

[Bre01] **L. Breiman** - *Random forests*. J. Mach. Learn. 2001: 45: 5–32.

[Bun92] **W. Buntine** - *Learning classification trees*. Statistics and Computing,1992: 2: 63–73.

[B+11] **M. Berks, Z. Chen, S. Astley, and C. Taylor** - *Detecting and classifying linear structures in mammograms using random forests*, IPMI 2011: LNCS 6801, pp. 510–524.

[Car09] **A.M. Carvalho** - *Scoring function for learning bayesian networks, Technical report, INESC-ID Tec*. Rep.2009: 54.

[CH92] **G. Cooper, E. Herskovitz** - *A Bayesian method for the induction of probabilistic networks from data*, Machine Learning, 1992: 9: 309-347.

[CMH03] **D.M. Chickering, C. Meek, D. Heckerman** - *Large-sample learning of Bayesian networks is NP-hard*, in: Proceedings of the 19th Conference on Uncertainty in Artificial Intelligence, Morgan-Kaufmann, 2003: pp. 124–133.

[CSG04] **R. Cano, C. Sordo, J.M. Gutie´rrez** - *Applications of Bayesian networks in meteorology*, Advances in Bayesian

Networks, in: J.A. Ga´mez et al. (Eds.), Springer-Verlag, pp. 309–327, 2004.

[DeG70] **M.H. DeGroot** - *Optimal Statistical Decision*, McGraw-Hill, 1970.

[Die00] **T.G. Dietterich** - *An experimental comparison of three methods for constructing ensembles of decision trees: bagging, boosting, and randomization*. Machine Learning, 2000: 40(2): 139–157.

[E+07] **R. Estevam, J. Hruschka, F.F. Nelson and Ebecken** - *Towards efficient variables ordering for Bayesian networks classifier*, Data & Knowledge Engineering, 2007: 63, 258–269.

[FS95] **Y. Freund, R. Schapire** - *A decision-theoretic generalization of on-line learning and an application to boosting*. In: Computational Learning Theory: Proceedings of the Second European Conference, 1995: pp. 23–37.

[F+00] **N. Friedman, M. Linial, I. Nachman, D. Peter** - *Using Bayesian networks to analyze expression data*, in: Proceedings of the fourth international annual conference on computational molecular biology, ACM Press, 2000: pp. 127–135.

[Hal00] **M. A. Hall** - *Correlation-Based Feature Selection for Discrete and Numeric Class Machine Learning*, Proc. 17th Int'l Conf. Machine Learning (ICML2000), 2000.

[Hsu04] **W.H. Hsu** - *Genetic wrappers for feature selection in decision tree induction and variable ordering in Bayesian network structure learning*, Information Science, 2004: 163: 103–122.

[H+09] **M. A. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, I. H. Witten** - *The WEKA data mining software: an update*, ACM SIGKDD Explorations, 2009: Volume 11, Issue 1.

[KP08] **Ł. Kobyliński, A. Przepiórkowski** - *Definition extraction with balanced random forests*, GoTAL 2008, LNAI, 2008: 5221, pp. 237–247.

[M+11] **B.H. Menze, B.M. Kelm, D.N. Splitthoff, U. Koethe and F.A. Hamprecht** - *On Oblique Random Forests*, ECML PKDD 2011: Part II, LNAI 6912, pp. 453–469.

[Ozc12] **A. Ozcift** - *Enhanced cancer recognition system based on random forests feature elimination algorithm*, J. Med Syst, 2012: 36:2577–2585.

[OPS11] **J. O. Ogutu, H.P. Piepho, T. Schulz-Streeck** - *A comparison of random forests, boosting and support vector machines for genomic selection*, BMC Proceedings, 2011: 5 (Suppl 3): S11.

[Qui93] **J.R. Quinlan** - *C4.5: Programs for machine learning*. Morgan Kaufmann, San Mateo, 1993.

[Rob04] **M. Robnik-Šikonja** - *Improving random forests*. In: J.F. Boulicaut et al. (eds.), Proc. 15th European Conf. on Machine Learning ECML 04, Springer, LNCS 3201, 359-370, 2004.

[TC09] **Thomas H. Cormen** - *Introduction to Algorithms* (3rd ed.). Massachusetts Institute of Technology, 2009: pp. 253–280. ISBN 978-0-262-03384-8.

[Vap99] **V. Vapnik** - *The nature of Statistical Learning theory*. Springer, Heidelberg, 1999.

[YDH08] **B-S. Yang, X. Di, T. Han** - *Random forests classifier for machine fault diagnosis*, Journal of Mechanical Science and Technology, 2008: 22: 1716-1725.

[Z+04] **M. D. Zio, M. Scanu, L. Coppola, O. Luzi, and A. Ponti** - *Bayesian networks for imputation*, Journal of the Royal Statistical Society A, 2004: 167 (Part 2) 309–322.