Correlation-Based and Causal Feature Selection Analysis for Ensemble Classifiers

Rakkrit Duangsoithong and Terry Windeatt

Center for Vision, Speech and Signal Processing
University of Surrey
Guildford, United Kingdom GU2 7XH
{r.duangsoithong,t.windeatt}@surrey.ac.uk

Abstract. High dimensional feature spaces with relatively few samples usually leads to poor classifier performance for machine learning, neural networks and data mining systems. This paper presents a comparison analysis between correlation-based and causal feature selection for ensemble classifiers. MLP and SVM are used as base classifier and compared with Naive Bayes and Decision Tree. According to the results, correlation-based feature selection algorithm can eliminate more redundant and irrelevant features, provides slightly better accuracy and less complexity than causal feature selection. Ensemble using Bagging algorithm can improve accuracy in both correlation-based and causal feature selection.

Keywords: Correlation-based feature selection, causal feature selection, ensemble classification.

1 Introduction

With improvements in information and technology, many information databases have been created. However, in some applications especially in biomedical area, dataset usually contains hundreds to thousands of features with relatively small sample size and leads to degradation in accuracy and efficiency of system by curse of dimensionality and over-fitting. The resulting classifier works very well with training data but very poorly on testing data.

To overcome this high dimensional feature spaces degradation problem, number of features should be reduced. Basically, there are two methods to reduce the dimension: feature extraction and feature selection. Feature extraction transforms or projects original features to fewer dimensions without using prior knowledge. Nevertheless, it lacks comprehensibility and uses all original features which may be impractical in large feature spaces. On the other hand, feature selection selects optimal feature subsets from original features by removing irrelevant and redundant features. It has the ability to reduce over-fitting, increase classification accuracy, reduce complexity, speed of computation and improve comprehensibility by preserving original semantic of datasets. Normally, clinicians prefer feature selection because of its understandbility and user acceptance.

F. Schwenker and N. El Gayar (Eds.): ANNPR 2010, LNAI 5998, pp. 25–36, 2010.

Feature selection is an important pre-processing step whether the classifier is Multilayer Perceptron (MLP), Support Vector Machines (SVM) or any other classifier. Generally, feature selection can be divided into four categories: Filter, Wrapper, Hybrid and Embedded methods [1],[2],[3]. Filter method is independent from learning method used in the classification process and uses measurement techniques such as correlation, distance and consistency measurement to find a good subset from entire set of features. Nevertheless, the selected subset may or may not be appropriate with the learning method. Wrapper method uses pre-determined learning algorithm to evaluate selected feature subsets that are optimum for the learning process. This method has high accuracy but is computationally expensive. Hybrid method combines advantage of both Filter and Wrapper method together. It evaluates features by using an independent measure to find the best subset and then uses a learning algorithm to find the final best subset. Finally, Embedded method interacts with learning algorithm but it is more efficient than Wrapper method because the filter algorithm has been built with the classifier.

Basically, feature selection does not take causal discovery or casuality into account [4]. Nevertheless, in some cases such as when training and testing dataset do not conform to i.i.d. assumption, testing distribution is shifted from manipulation by external agent, causal discovery can provide some benefits for feature selection under these uncertainty conditions. Causality also can learn underlying data structure, provide better understanding of the data generation process and better accuracy and robustness under uncertainty conditions [4].

Normally, causal relationships are uncovered by Bayesian Networks (BNs) which consists of a direct acyclic graph (DAG) that represents dependencies and independencies between variable and joint probability distribution among a set of variables [5].

An ensemble classifier or multiple classifier system (MCS) is another well-known technique to improve system accuracy [6]. Ensemble combines multiple base classifiers to learn a target function and gathers their prediction together. It has ability to increase accuracy of system by combining output of multiple experts to reduce bias and variance, improve efficiency by decomposing complex problem into multiple sub problems and improve reliability by reducing uncertainty. To increase accuracy, each classifier in the ensemble should be diverse or unique in order to reduce total error such as starting with different input, initial weight, random features or random classes [7].

In this paper, we present a comparison analysis between correlation-based and causal feature selection for ensemble classifiers in terms of number of eliminated features, complexity of algorithms and average percent accuracy.

1.1 Related Research

Feature selection and ensemble classification have received attention from many researchers in statistics, machine learning, neural networks and data mining areas for many years. At the beginning of feature selection history, most researchers focused only on removing irrelevant features such as ReliefF [8], FOCUS [9] and

Correlation-based Feature Selection(CFS) [10]. Recently, in Yu and Liu (2004) [11], Fast Correlation-Based Filter (FCBF) algorithm was proposed to remove both irrelevant and redundant features by using Symmetrical Uncertainty (SU) measurement and was successful for reducing high dimensional features while maintaining high accuracy.

In the past few years, learning BNs from observation data has received increasing attention from researchers for many applications such as decision support system, information retrieval, natural language processing, feature selection and gene expression data analysis [12],[13].

The category of BNs can be divided into three approaches: Search-and-Score, Constraint-Based and Hybrid approaches [12],[13]. In Search-and-Score approach, BNs search all possible structures to find the one that provides the maximum score. The second approach, Constraint-Based, uses test of conditional dependencies and independencies from the data by estimation using G^2 statistic test or mutual information, etc. This approach defines structure and orientation from results of the tests based on some assumptions that these tests are accurate. Finally, Hybrid approach uses Constraint-Based approach for conditional independence test (CI test) and then identifies the network that maximizes a scoring function by using Search-and-Score approach [13].

Constraint-Based algorithms are computationally effective and suitable for high dimensional feature spaces. PC algorithm [14], is a pioneer, prototype and well-known global algorithm of Constraint-Based approach for causal discovery. Three Phase Dependency Analysis (TPDA or PowerConstructor) [15] is another global Constraint-Based algorithm that uses mutual information to search and test for CI test instead of using G^2 Statistics test as in PC algorithm. However, both PC and TPDA algorithm use global search to learn from the complete network that can not scale up to more than few hundred features (they can deal with 100 and 255 features for PC and TPDA, respectively) [16]. Sparse Candidate algorithm (SC) [17] is one of the prototype BNs algorithm that can deal with several hundreds of features by using locally candidate set. Nevertheless, SC algorithm has some disadvantages, it may not identify true set of parents and users have to find appropriate k parameter of SC algorithm by themselves [12].

Recently, many Markov Blanket-based algorithms for causal discovery have been studied extensively and they have ability to deal with high dimensional feature spaces such as MMMB, IAMB [16] and HITON [5] algorithms. HITON is a state-of-the-art algorithm that has ability to deal with thousands of features and can be used as an effective feature selection in high dimensional spaces. However, HITON and all other MB-based algorithms may not specify features in Markov Blanket for desired classes or target (MB(T)) when the data is not faithful [20].

2 Theoretical Approach

In our research, two correlation-based feature selection methods, Fast Correlation-Based Filter (FCBF) [11] and Correlation-based Feature Selection with Sequential Forward Floating Search (CFS+SFFS) [10],[18] are compared with causal

feature selection algorithms (PC, TPDA, SC and HITON) for Bagging [21] ensemble classifiers (described in Section 2.2) and experimentally compared with different learning algorithms.

2.1 Feature Selection

Fast Correlation-Based Filter (FCBF). FCBF [11] algorithm has two stages: relevance analysis and redundancy analysis.

Relevance Analysis. Normally, correlation is widely used to analyze relevance. In linear systems, correlation can be measured by linear correlation coefficient.

$$r = \frac{\sum_i (x_i - \overline{x_i})(y_i - \overline{y_i})}{\sqrt{\sum_i (x_i - \overline{x_i})^2} \sqrt{\sum_i (y_i - \overline{y_i})^2}} \tag{1}$$

However, most systems in real world applications are non-linear. Correlation in non-linear systems can be measured by using Symmetrical Uncertainty (SU).

$$SU = 2\Big[\frac{IG(X|Y)}{H(X)H(Y)}\Big] \tag{2}$$

$$IG(X,Y) = H(X) - H(X|Y) \tag{3}$$

$$H(X) = -\sum_i P(x_i) log_2 P(x_i) \tag{4}$$

where $IG(X|Y)$ is the Information Gain of X after observing variable Y. $H(X)$ and $H(Y)$ are the entropy of variable X and Y, respectively. $P(x_i)$ is the probability of variable x.

SU is the modified version of Information Gain that has range between 0 and 1. FCBF removes irrelevant features by ranking correlation (SU) between feature and class. If SU between feature and class equal to 1, it means that this feature is completely related to that class. On the other hand, if SU is equal to 0, the features are irrelevant to this class.

Redundancy Analysis. After ranking relevant features, FCBF eliminates redundant features from selected features based on SU between feature and class and between feature and feature. Redundant features can be defined from meaning of predominant feature and approximate Markov Blanket. In Yu and Liu (2004) [11], a feature is predominant (both relevant and non redundant feature) if it does not have any approximate Markov blanket in the current set.

Approximate Markov blanket: For two relevant features F_i and F_j $(i \neq j)$, F_j forms an approximate Markov blanket for F_i if

$$SU_{j,c} \geq SU_{i,c} \; and \; SU_{i,j} \geq SU_{i,c} \tag{5}$$

where $SU_{i,c}$ is a correlation between any feature and the class. $SU_{i,j}$ is a correlation between any pair of feature F_i and F_j $(i \neq j)$.

Correlation-based Feature Selection (CFS). CFS [10] is one of well-known techniques to rank the relevance of features by measuring correlation between features and classes and between features and other features.

Given number of features k and classes C, CFS defined relevance of features subset by using Pearson's correlation equation

$$Merit_s = \frac{kr_{kc}}{\sqrt{k + (k-1)r_{kk}}} \tag{6}$$

where $Merit_s$ is relevance of feature subset, r_{kc} is the average linear correlation coefficient between these features and classes and r_{kk} is the average linear correlation coefficient between different features.

Normally, CFS adds (forward selection) or deletes (backward selection) one feature at a time, however, in this research, we used Sequential Forward Floating Search (SFFS) as the search direction.

Sequential Forward Floating Search (SFFS). SFFS [18] is one of a classic heuristic searching method. It is a variation of bidirectional search and sequential forward search (SFS) that has dominant direction on forward search. SFFS removes features (backward elimination) after adding features (forward selection). The number of forward and backward step is not fixed but dynamically controlled depending on the criterion of the selected subset and therefore, no parameter setting is required.

Causal Discovery Algorithm. In this paper, three standard (PC, TPDA, SC) and one state-of-the-art causal discovery algorithms (HITON) are used as causal feature selection methods. In the final output of the causal graph from each algorithm, the unconnected features to classes will be considered as eliminated features.

1. PC Algorithm

PC algorithm [14],[4] is the prototype of constraint-based algorithm. It consists of two phases: Edge Detection and Edge Orientation.

Edge Detection: the algorithm determines directed edge by using conditionally independent condition. The algorithm starts with:

i) Undirected edge with fully connected graph.

ii) Remove a share direct edge between A and B ($A - B$) iff there is a subset F of features that can present conditional independence $(A, B|F)$.

Edge Orientation: The algorithm discovers V-Structure $A - B - C$ in which $A - C$ is missing.

i) If there are direct edges between $A - B$ and $B - C$ but not $A - C$, then orient edge $A \rightarrow B \leftarrow C$ until no more possible orientation.

ii) If there is a path $A \rightarrow B - C$ and $A - C$ is missing, then $A \rightarrow B \rightarrow C$.

iii) If there is orientation $A \rightarrow B \rightarrow ... \rightarrow C$ and $A - C$ then orient $A \rightarrow C$.

2. Three Phase Dependency Analysis Algorithm (TPDA)
TPDA or PowerConstructor algorithm [15] has three phases: *drafting, thickening and thinning*. In *drafting phase*, mutual information of each pair of nodes is calculated and used to create a graph without loop. After that, in *thickening phase*, edge will be added when that pair of nodes can not be *d-separated*. (node A and B are *d-separated* by node C iff node C blocks every path from node A to node B [12].) The output of this phase is called an independence map (*I-map*). The edge of *I-map* will be removed in *thinning phase* if two nodes of the edge can be *d-separated* and the final output is defined as a *perfect map* [15].

3. Sparse Candidate Algorithm (SC)
SC algorithm has two phases: *restrict* and *maximize steps* [17]. In *restrict step*, candidate sets are chosen by heuristic estimates of size k (define by user) and then a hill climbing search is performed in *maximize step*. In this second step, a network is started with empty graph and one of the operators: *add*, *delete* or *reverse* that provides the highest score will be chosen and applied to the current network. Finally, the algorithm will be repeated until there is no change in the candidate set [17],[12],[19].

4. HITON Algorithm
HITON [5] is one of state-of-the-art causal discovery algorithms that can be used as feature selection and can scale up to deal with thousands of features. HITON identifies Markov Blanket of the classes (or target) and then removes by backward greedy wrapper search of the features from the Markov Blanket that do not affect the classifier performance [5],[20].

2.2 Ensemble Classifier

Bagging. Bagging [21] or Bootstrap aggregating is one of the earliest, simplest and most popular for ensemble based classifiers. Bagging uses Bootstrap that randomly samples with replacement and combines with majority vote. Bootstrap is the most well-known strategy for injecting randomness to improve generalization performance in multiple classifier systems and provides out-of-bootstrap estimate for selecting classifier parameters [6]. Randomness is desirable since it increases diversity among the base classifiers, which is known to be a necessary condition for improved performance. However, there is an inevitable trade off between accuracy and diversity known as the accuracy/diversity dilemma [6].

3 Experimental Setup

3.1 Dataset

The medical datasets used in this experiment were taken from UCI machine learning repository [22]: heart disease, hepatitis, diabetes and Parkinson dataset and from Causality Challange [23]: lucas and lucap datasets. The details of

Table 1. Datasets

Dataset	Sample	Features	Classes	Missing Values	Data type
Heart Disease	303	13	5	Yes	Numeric (cont. and discrete)
Diabetes	768	8	2	No	Numeric (continuous)
Hepatitis	155	19	2	Yes	Numeric (cont. and discrete)
Parkinson's	195	22	2	No	Numeric (continuous)
Lucas	2000	11	2	No	Numeric (binary)
Lucap	2000	143	2	No	Numeric (binary)

datasets are shown in Table 1. The missing data are replaced by mean and mode of that dataset.

3.2 Evaluation

To evaluate feature selection process we use four widely used classifiers: Naive-Bayes(NB), Multilayer Perceptron (MLP), Support Vector Machines (SVM) and Decision Trees (DT). The parameters of each classifier were chosen based on the highest average accuracy of the experiment datasets from base classifier. MLP has one hidden layer with 16 hidden nodes, learning rate 0.2, momentum 0.3, 500 iterations and uses backpropagation algorithm with sigmoid transfer function. SVMs uses linear kernel and set the regularization value to 0.7 and Decision Trees use pruned C4.5 algorithm. The number of classifiers in Bagging is varied from 1, 5, 10, 25 to 50 classifiers. The threshold value of FCBF algorithm in our research is set at zero for heart disease, diabetes, parkinson and lucas and 1.4 and 0.15 for hepatitis and lucap dataset, respectively.

The classifier results were validated by 10 fold cross validation with 10 repetitions for each experiment and evaluated by average percent of test set accuracy of algorithm. For causal feature selection, PC algorithm using mutual information as statistic test with threshold 0.01 and maximum cadinality equals to 2. In TPDA algorithm, mutual information are used as statistic test with threshold 0.01 and assumed that data is monotone faithful. SC algorithm uses BDeu score function, k = 5 and using Bayesian scoring metric for statistic test. Finally, HITON use G^2 statistic test with threshold 0.05, maximum size of the conditional set is set to 3 and provides output as Markov Blanket of the classes.

4 Experimental Result

Table 2 and table 3 show the number of selected features in each analysis and the complexity of the algorithm, respectively.

Figure 1 and 2 present example of the average accuracy of heart disease and lucas dataset. Y-axis presents the average percent accuracy of the classifier and X-axis shows the number of ensemble from 1 to 50 classifiers. Figure 3 and 4 show the average accuracy of six datasets for each classifier and average of all classifiers for all six datasets, respectively. Finally, figure 5 presents the examples of causal graph of lucas data set from PC algorithm.

Table 2. Number of selected features

Dataset	Original Feature	Correlation-Based		Causal			
		FCBF	CFS+SFFS	PC	TPDA	SC	HITON
Heart Disease	13	6	9	13	13	11	4
Diabetes	8	4	4	8	8	0	0
Hepatitis	19	3	10	19	18	0	0
Parkinson's	22	5	10	22	2	0	0
Lucas	11	3	3	9	10	11	0
Lucap	143	7	36	121	121	123	0

Table 3. The complexity of each algorithm

Algorithm	Complexity	Remark
FCBF	$O(MNlogN)$	M=number of samples, N= number of features
$CFS + SFFS$	$< O(N^2)$	N= number of features
PC	$O(\|N\|^4)$	N= number of features
TPDA	$O(\|N\|^4)$	N= number of features
SC	$O(2^k \cdot (c+1)! \cdot \|J\|)$	k = size of candidate set, c = size of the largest separator in cluster tree, J = a family of cluster
HITON	$O(\|MB\|^3 \|N\|)$	MB = Markov Blanket of the class, N = number of features

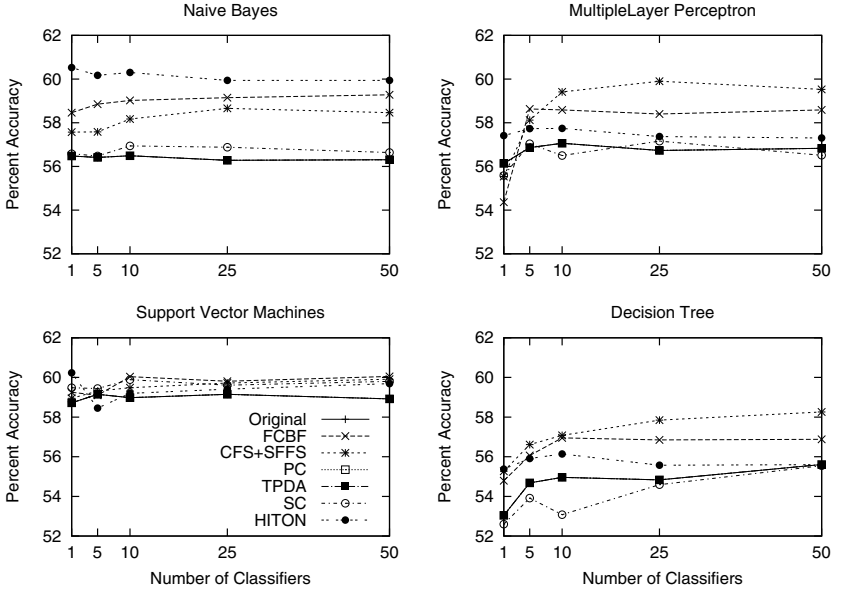

Fig. 1. Average Percent Accuracy of Heart Disease dataset

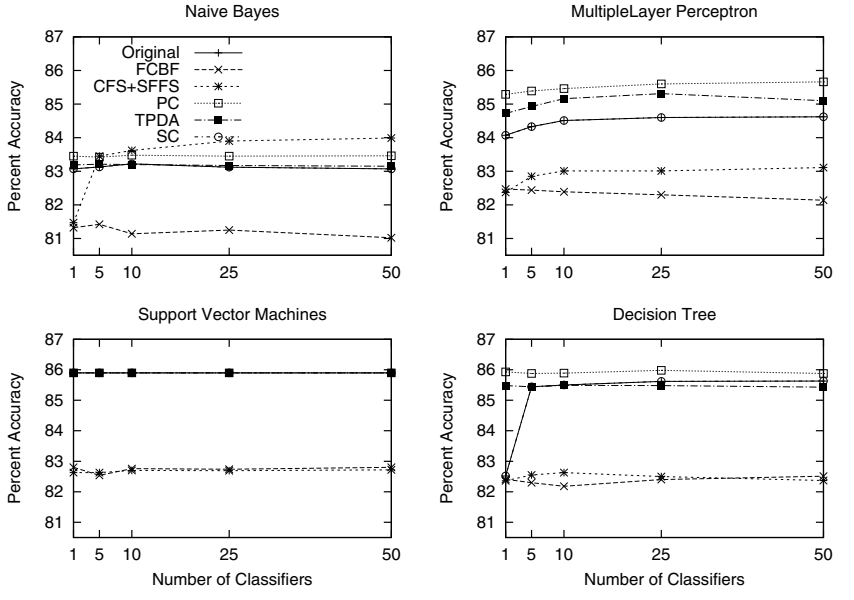

Fig. 2. Average Percent Accuracy of Lucas dataset

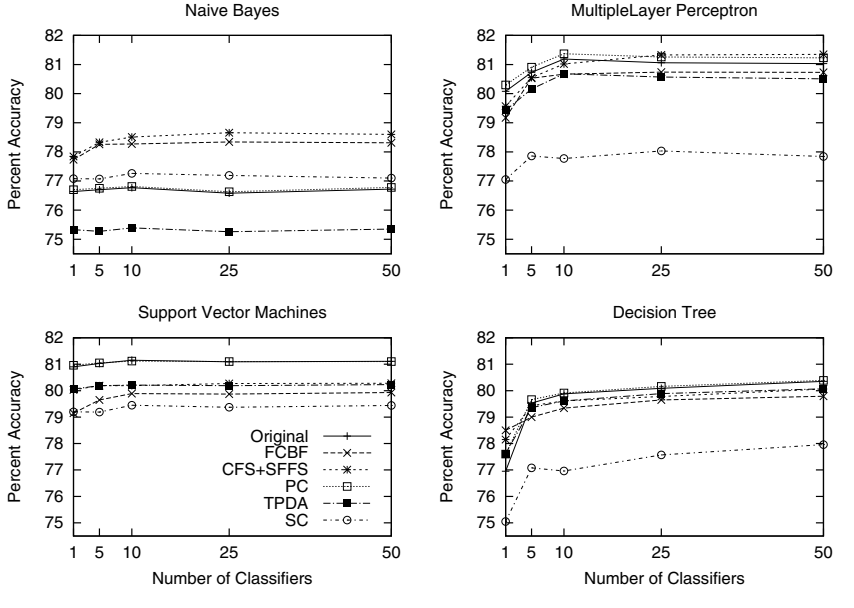

Fig. 3. Average Percent Accuracy of six datasets for each classifier

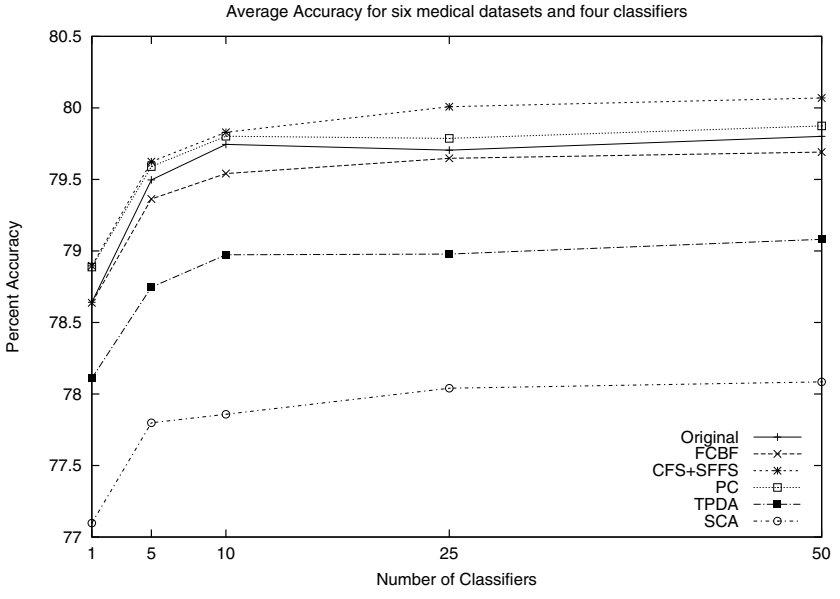

Fig. 4. Average Percent Accuracy of six datasets, four classifiers

5 Discussion

According to table 2, it can be seen that HITON eliminates highest number of features among other algorithms, however, it can define Markov Blanket for only heart disease and does not find any Markov Blanket for the remaining datasets because the data distribution may not be faithful [20] . SC algorithm also eliminates all features in some datasets because it may not identify true set of parents when k parameter is not appropriate [12]. FCBF algorithm removes more features than CFS+SFFS, TPDA and PC algorithms, respectively.

From Table 3, CFS+SFFS has the least complexity among other algorithms. HITON does not have high complexity because it uses Markov Blanket discovery that select only parents, children and spouses of the classes. PC and TPDA have the highest complexity algorithm due to their exhaustive search.

With reference to figure 3, SVM provides less accuracy than MLP because MLP uses back propagation with sigmoid transfer function and has 16 hidden node which is non linear system while SVM uses linear kernel with regularization 0.7. In figure 4, CFS+SFFS provides better average accuracy than PC, original, FCBF, TPDA and SCA, respectively. (HITON algorithm which can select optimal features only in heart disease dataset is not considered in the average graph in order not to bias result.) PC gives the best average accuacy among causal feature selection algorithms, however, it can deal only with few hundred features [16]. FCBF does not provide highest accuracy because its main objective is dealing with high dimensional feature while preserving high accuracy [11].

Although causal feature selection provides slightly less accuracy, more complexity and less number of eliminated features than correlation-based feature selection, it has benefit to learn underlying causal structure of the classes and features as an example shown in figure 5.

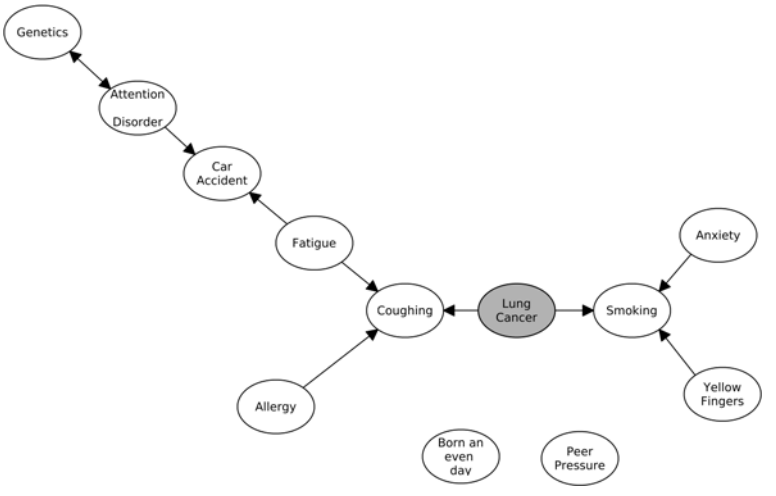

Fig. 5. Causal structure of Lucas dataset from PC algorithm

6 Conclusion

In this paper, we present a comparison analysis between correlation-based and casual feature selection for ensemble classifiers. In conclusion, correlation-based feature selection has slightly higher average accuracy, less complexity and can remove more irrelevant and redundant features than causal feature selection. Nevertheless, causal feature selection can reveal causes and consequence of the classes by defining causal relationship. Ensemble has ability to improve both correlation-based and causal feature selection. The future work will examine the result of causal feature selection from bootstrap dataset and combine result with ensemble classifiers.

References

1. Liu, H., Yu, L.: Toward integrating feature selection algorithms for classification and clustering. IEEE Transactions on Knowledge and Data Engineering 17(4), 491–502 (2005)
2. Saeys, Y., Inza, I., Larranaga, P.: A review of feature selection techniques in bioinformatics. Bioinformatics 23(19), 2507–2517 (2007)
3. Duangsoithong, R., Windeatt, T.: Relevance and Redundancy Analysis for Ensemble Classifiers. In: Perner, P. (ed.) Machine Learning and Data Mining in Pattern Recognition. LNCS, vol. 5632, pp. 206–220. Springer, Heidelberg (2009)

4. Guyon, I., Aliferis, C., Elisseeff, A.: Causal Feature Selection. In: Liu, H., Motoda, H. (eds.) Computational Methods of Feature Selection. Chapman and Hall, Boca Raton (2007)
5. Aliferis, C.F., Tsamardinos, I., Statnikov, A.: HITON, A Novel Markov Blanket Algorithm for Optimal Variable Selection. In: AMIA 2003 Annual Symposium Proceedings, pp. 21–25 (2003)
6. Windeatt, T.: Ensemble MLP Classifier Design, vol. 137, pp. 133–147. Springer, Heidelberg (2008)
7. Windeatt, T.: Accuracy/diversity and ensemble MLP classifier design. IEEE Transactions on Neural Networks 17(5), 1194–1211 (2006)
8. Witten, I.H., Frank, E.: Data Mining: Practical Machine Learning Tools and Techniques, 2nd edn. Morgan Kaufmann, San Francisco (2005)
9. Almuallim, H., Dietterich, T.G.: Learning with many irrelevant features. In: Proceedings of the Ninth National Conference on Artificial Intelligence, pp. 547–552. AAAI Press, Menlo Park (1991)
10. Hall, M.A.: Correlation-based feature selection for discrete and numeric class machine learning. In: Proceeding of the 17th International Conference on Machine Learning, pp. 359–366. Morgan Kaufmann, San Francisco (2000)
11. Yu, L., Liu, H.: Efficient feature selection via analysis of relevance and redundancy. J. Mach. Learn. Res. 5, 1205–1224 (2004)
12. Tsamardinos, I., Brown, L.E., Aliferis, C.F.: The max-min hill-climbing Bayesian network structure learning algorithm. Machine Learning 65, 31–78 (2006)
13. Wang, M., Chen, Z., Cloutier, S.: A hybrid Bayesian network learning method for constructing gene networks. Computational Biology and Chemistry 31, 361–372 (2007)
14. Spirtes, P., Glymour, C., Schinese, R.: Causation, Prediction, and search. Springer, New York (1993)
15. Cheng, J., Bell, D., Liu, W.: Learning Belief Networks from Data: An Information theory Based Approach. In: Proceedings of the Sixth ACM International Conference on Information and Knowledge Management, pp. 325–331 (1997)
16. Tsamardinos, I., Aliferis, C.F., Statnikov, A.: Time and Sample Efficient Discovery of Markov Blankets and Direct Causal Relations. In: KDD 2003, Washington DC, USA (2004)
17. Friedman, N., Nachman, I., Peer, D.: Learning of Bayesian Network Structure from Massive Datasets: The "Sparse Candidate" Algorithm. In: Proceedings of the 15th Conference on Uncertainty in Artificial Intelligence (UAI), pp. 206–215. Morgan Kaufmann, Stockholme (1999)
18. Pudil, P., Novovicova, J., Kitler, J.: Floating Search Methods in Feature Selection. Pattern Recognition Letters 15, 1,119–1,125 (1994)
19. Brown, L.E., Tsamardinos, I., Aliferis, C.F.: A Novel Algorithm for Scalable and Accurate Bayesian Network Learning. Medinfo. 11, 711–715 (2004)
20. Brown, L.E., Tsamardinos, I.: Markov Blanket-Based Variable Selection. Technical Report DSL TR-08-01 (2008)
21. Breiman, L.: Bagging predictors. Machine Learning 24(2), 123–140 (1996)
22. Asuncion, A., Newman, D.: UCI machine learning repository (2007), `http://www.ics.uci.edu/mlearn/MLRepository.html`
23. Guyon, I.: Causality Workbench (2008), `http://www.causality.inf.ethz.ch/home.php`