# Hybrid Correlation and Causal Feature Selection for Ensemble Classifiers

Rakkrit Duangsoithong and Terry Windeatt

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

**Abstract.** PC and TPDA algorithms are robust and well known prototype algorithms, incorporating constraint-based approaches for causal discovery. However, both algorithms cannot scale up to deal with high dimensional data, that is more than few hundred features. This paper presents hybrid correlation and causal feature selection for ensemble classifiers to deal with this problem. The number of eliminated features, accuracy, the area under the receiver operating characteristic curve (AUC) and false negative rate (FNR) of proposed algorithms are compared with correlation-based feature selection (FCBF and CFS) and causal based feature selection algorithms (PC, TPDA, GS, IAMB).

**Key words:** Feature selection, causal discovery, ensemble classification.

## 1 Introduction

Feature selection is an important pre-processing step to reduce feature dimensions for classification and generally, can be divided into four categories [1],[2],[3]. Filter method is independent from learning method and uses measurement techniques such as correlation and distance measurement to find a good subset from entire set of features. Wrapper method uses pre-determined learning algorithm to evaluate selected feature subsets that are optimum for the learning process. 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.

Feature selection does not usually take causal discovery into account. However, 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 [4].

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

Generally, the category of BNs can be divided into: Search-and-Score and Constraint-Based approaches. 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 (CI) from the data by estimation using $G^2$ statistic test or mutual information, etc. Constraint-Based algorithms are computationally effective and suitable for high dimensional feature spaces. PC algorithm [6], is a pioneer, prototype and well-known global algorithm of Constraint-Based approach for causal discovery. Three Phase Dependency Analysis (TPDA or PowerConstructor) [7] 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 and can not scale up to more than few hundred features (they can deal with 100 and 255 features for PC and TPDA, respectively) [8]. 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 GS [9], MMMB, IAMB [8] and HITON [5] algorithms.

An ensemble classifier or multiple classifier system (MCS) is another well-known technique to improve system accuracy [10]. Ensemble combines multiple base classifiers to learn a target function and gathers their prediction together. It has ability to increase accuracy 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 such as starting with different input, initial weight, random features or random classes [11].

The main objective of this paper is to find algorithm that can scale up PC and TPDA algorithms to deal with high dimensional data. We propose analysis of hybrid correlation and causal feature selection for ensemble classifiers in terms of number of eliminated features, average percent accuracy, the area under the receiver operating characteristic curve (AUC) and false negative rate (FNR).

## 2 Theoretical Approach

In our research, hybrid algorithm of correlation and causal feature selection is compared with Fast Correlation-Based Filter (FCBF), Correlation-based Feature Selection with Sequential Forward Floating Search direction (CFS+SFFS), and with causal feature selection algorithms (PC, TPDA, GS and IAMB) using Bagging (described in Section 2.2).

## 2.1 Feature Selection

**2.1.1 Correlation-based Redundancy and Relevance Analysis** The concept of selecting optimal subset from whole features is presented in Figure 1 [12]. where I is irrelevant feature, II is weakly relevant and redundant feature, III is weakly relevant but non redundant feature. IV is strongly relevant feature and III+IV are optimal subset.

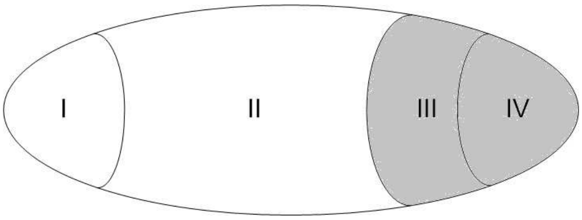

**Fig. 1.** Optimal Subset

Optimal subset should include all strongly relevant features, subset of weakly relevant features that have no redundancy and none of the irrelevant features.

Table 1 shows the summary analysis of redundancy and relevancy analysis for correlation-based [12], causal-based [4] and proposed hybrid correlation and causal feature selection. Markov Blanket (MB(T)) of target or class (T) is the minimal set of conditional features that all other features are probabilistically independent of T. It consists of the set of parents, children and spouses of T. Approximate Markov Blanket is explained section 2.1.1 a.

**Table 1.** Summary analysis of correlation, causal and proposed hybrid correlation and causal feature selection for redundancy and relevance analysis.

| Relation | Correlation-Based | Causal-Based | Hybrid algorithm |
|---|---|---|---|
| Strongly relevant | $SU_{i,c} = 1$ | Features in Markov Blanket | Features in Markov Blanket |
| Weakly relevant without redundant features | does not has approximate Markov Blanket | connected to classes | connected to classes |
| Weakly relevant with redundant features | has approximate Markov Blanket | connected to classes | has approximate Markov Blanket |
| Irrelevant | $SU_{i,c} = 0$ | disconnected to classes | disconnected to classes |

**a) Fast Correlation-Based Filter (FCBF).** FCBF [12] algorithm has two stages: relevance analysis and redundancy analysis.

*Relevance Analysis.* Correlation can be measured by using Symmetrical Uncertainty (SU).

$$SU(X,Y) = 2\Big[\frac{IG(X|Y)}{H(X)+H(Y)}\Big] \tag{1}$$

$$IG(X|Y) = H(X) - H(X|Y) \tag{2}$$

$$H(X) = -\sum_i P(x_i)log_2 P(x_i) \tag{3}$$

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.* Redundant features can be defined from meaning of predominant feature and approximate Markov Blanket. In Yu and Liu (2004) [12], 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

$$SU_{j,c} \geq SU_{i,c} \; and \; SU_{i,j} \geq SU_{i,c} \tag{4}$$

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 $F_j$ $(i \neq j)$.

**b) Correlation-based Feature Selection (CFS).** CFS [13] 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

$$Merit_s = \frac{kr_{kc}}{\sqrt{k+(k-1)r_{kk}}} \tag{5}$$

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) [14] as the search direction.

**2.1.2 Causal Discovery Algorithm.** In this paper, two standard constraint-based approaches (PC and TPDA) and two Markov Blanket based algorithms (GS, IAMB) 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.

**a) PC Algorithm** PC algorithm [6],[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 $A \rightarrow C$.

**b) Three Phase Dependency Analysis Algorithm (TPDA)** TPDA or PowerConstructor algorithm [7] 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 [15].) 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* [7].

**c) Grow-Shrink algorithm (GS)** GS [9] algorithm consists of two phases, forward and backward.

*Forward phase*: GS statistically ranks features by using the strength of association with target or class (T) given empty set. After that the next ordering feature which is not conditionally independent from class T given current Markov Blanket (CMB) will added into CMB.

*Backward phase*: Identify false positive nodes and remove them from CMB. At this stage, $CMB = MB(T)$. Finally, a feature X will be removed from CMB one-by-one if that feature X is independent of class T given the remaining CMB.

**d) Incremental Association Markov Blanket Algorithm. (IAMB)** IAMB [8] is one of Markov Blanket detection algorithms using forward selection followed by removing false positive node. IAMB has two phases, forward and backward.

*Forward phase*: In forward selection phase, the algorithm starts with empty set in CMB, then adding features which maximizes a heuristic function $f(X; T|CMB)$. A feature member in MB(T) will not return zero value of this function.

*Backward phase*: False positive nodes will be removed from CMB by using condition independent testing of class T given the rest CMB.

### 2.2 Ensemble Classifier

Bagging [16] or **B**ootstrap **agg**regat**ing** is one of the earliest, simplest and most popular methods for ensemble based classifiers. Bagging uses Bootstrap that randomly samples with replacement and combines with majority vote. The selected data is divided to $m$ bootstrap replicates and randomly sampled with replacement. Each bootstrap replicate contains, on average, 63.2 % of the original dataset. Final output will be selected from majority vote from all classifiers of each bootstrap replicate. 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 [10]. 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 [10].

## 3 Experimental Setup

### 3.1 Dataset

The datasets used in this experiment were taken from Causality Challenge [17] and details of each dataset are shown in Table 2.

**Table 2.** Datasets.

| Dataset | Sample | Features | Classes | Missing Values | Data type |
|---|---|---|---|---|---|
| LUCAS | 2000 | 11 | 2 | No | Numeric (binary) |
| LUCAP | 2000 | 143 | 2 | No | Numeric (binary) |
| REGED | 500 | 999 | 2 | No | Numeric (discrete) |
| CINA | 16033 | 132 | 2 | No | Numeric (discrete) |
| SIDO | 12678 | 4932 | 2 | No | Numeric (binary) |

### 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 as follows. 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. SVM uses polynomial kernel with exponent 2 and the regularization value set to 0.7. DT uses 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 LUCAS, REGED, CINA, SIDO and 0.14 for 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, FNR and AUC.

Due to large number of samples and limitation of computer memory during validation in CINA and SIDO datasets, the number of samples of both dataset are reduced to 10 percent (1603 and 1264 samples, respectively) from the original dataset.

For causal feature selection, PC algorithm uses mutual information ($MI$) as statistic test with threshold 0.01 and maximum cardinality equal to 2. In TPDA algorithm, mutual information was used as statistic test with threshold 0.01 and data assumed to be monotone faithful. GS and IAMB algorithm use $MI$ statistic test with significant 0.01 and provides output as Markov Blanket of the classes.

## 4 Experimental Result

Table 3 presents the number of selected features for correlation-based, causal based feature selection and proposed hybrid algorithm. It can be seen that PC and TPDA algorithms are impractical for high dimensional features due to their complexity. However, if redundant features are removed, the number of selected features will enable both algorithms to be practical as shown in proposed hybrid algorithm. Nevertheless, for some datasets such as REGED, TPDA algorithm might not be feasible because of many complex connections between nodes (features).

**Table 3.** Number of selected features from each algorithm.

| Dataset | Original | Correlation-Based | | Causal-Based | | | | Hybrid algorithm | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | Feature | FCBF | CFS | PC | TPDA | GS | IAMB | H-PC | H-TPDA | H-GS | H-IAMB |
| LUCAS | 11 | 3 | 3 | 9 | 10 | 9 | 11 | 2 | 3 | 2 | 2 |
| LUCAP | 143 | 7 | 36 | 121 | 121 | 16 | 14 | 21 | 22 | 17 | 13 |
| REGED | 999 | 18 | 18 | N/A | N/A | 2 | 2 | 18 | N/A | 2 | 2 |
| CINA | 132 | 10 | 15 | 132 | N/A | 4 | 4 | 5 | 7 | 10 | 9 |
| SIDO | 4932 | 23 | 25 | N/A | N/A | 17 | 17 | 2 | 3 | 1 | 2 |

Figure 2 to 4 show the average percent accuracy, AUC and FNR of five datasets for all classifiers. From average accuracy in figure 2, correlation-based feature selection (FCBF, CFS) provides the best average accuracy. Hybrid correlation and causal feature selection has better accuracy than original causal feature selection. Hybrid method using PC algorithm (H-PC) has slightly lower average accuracy than correlation-based feature selection but has the ability to deal with high dimensional features. From figure 3, PC, CFS, TPDA and FCBF algorithm provide the best and comparable AUC. Proposed hybrid algorithm has lower AUC than both correlation and original causal-based algorithms. In figure 4, H-PC has the lowest FNR. In all experiments, hybrid algorithm provides lower FNR than original causal algorithm but still higher than correlation-based algorithm.

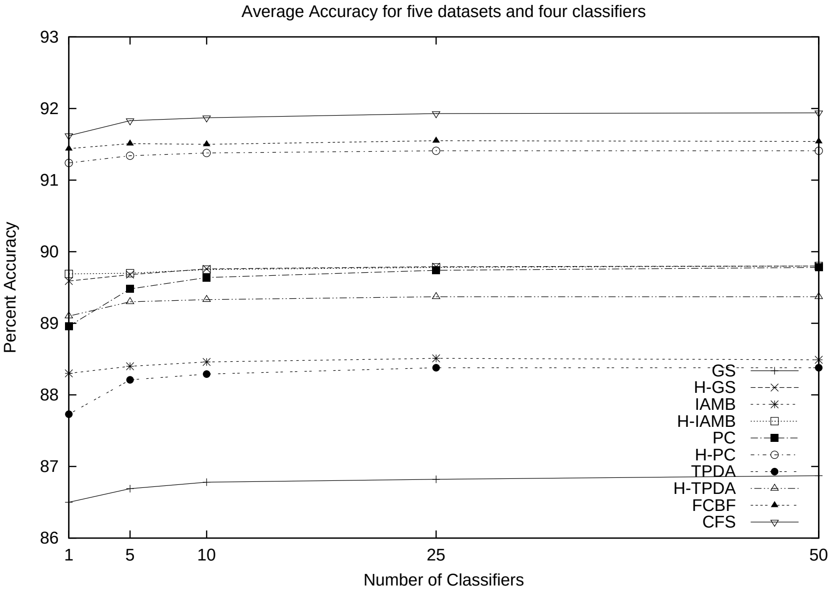

**Fig. 2.** Average Percent Accuracy of five datasets and four classifiers

Ensemble classifiers using Bagging slightly improves accuracy and AUC for most algorithms. Bagging also reduces FNR for CFS, PC and TPDA algorithm but provides stable FNR for the rest. After increasing number of classifiers to 5-10, the graphs of average accuracy, AUC and FNR all reach saturation point.

## 5 Conclusion

In this paper, hybrid correlation and causal feature selection for ensemble classifiers is presented to deal with high dimensional features. According to the results, the proposed hybrid algorithm provides slightly lower accuracy, AUC and higher FNR than correlation-based. However, compared to causal-based feature selection, the proposed hybrid algorithm has lower FNR, higher average accuracy and AUC than original causal-based feature selection. Moreover, the proposed hybrid algorithm can enable PC and TPDA algorithms to deal with high dimensional

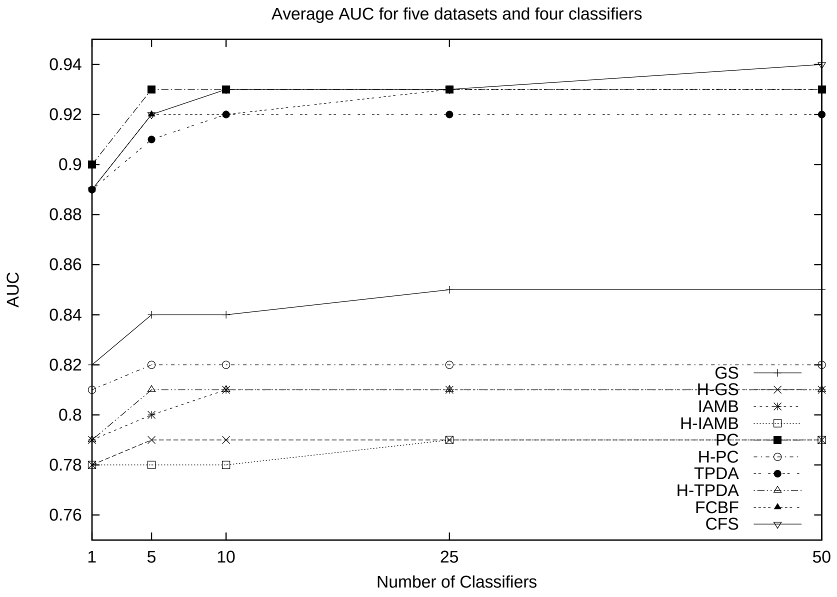

**Fig. 3.** Average AUC of five datasets and four classifiers

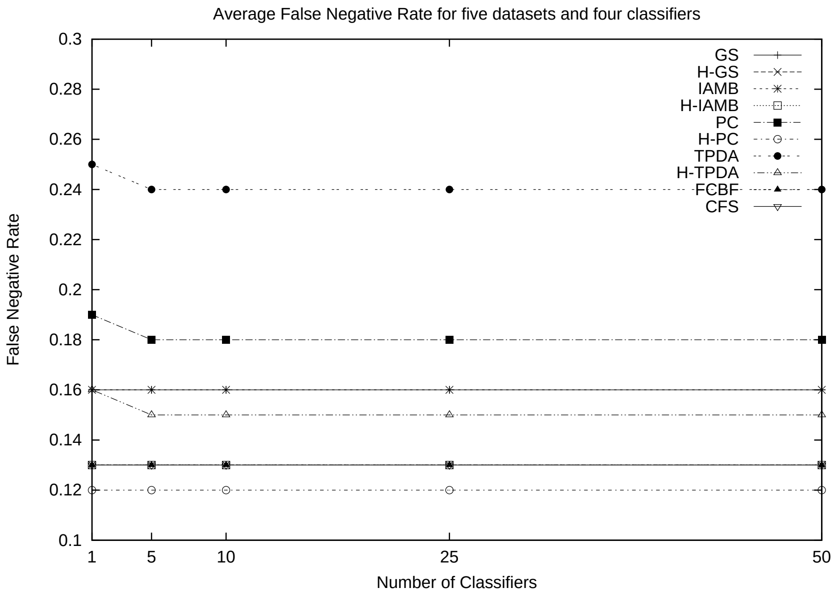

**Fig. 4.** Average FNR of five datasets and four classifiers

features while maintaining high accuracy, AUC and low FNR. Also the underlying causal structure is more understandable and has less complexity. Ensemble classifiers using Bagging provide slightly better results than single classifier for most algorithms. The future work will improve accuracy of search direction in structure learning for causal feature selection algorithm.

## 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, vol. 5362, pp. 206–220, Springer, Heidelberg (2009).
4. Guyon, I., Aliferis, C., Elisseeff, A.: Causal Feature Selection. In Computational Methods of Feature Selection, Liu, H. and Motoda, H. editors. Chapman and Hall (2007)
5. Aliferis, C.F., Tsamardinos, I., Statnikov, A.: HITON, A Novel Markov Blanket Algorithm for Optimal Variable Selection. AMIA 2003 Annual Symposium Proceedings, pp 21–25.(2003)
6. Spirtes, P. Glymour, C., Schinese, R.: Causation, Prediction, and search. Springer, New York (1993)
7. Cheng, J., Bell, D.A., 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)
8. Tsamardinos, I., Aliferis, C.F., Statnikov,A.: Time and Sample Efficient Discovery of Markov Blankets and Direct Causal Relations. KDD2003 Washington DC, USA. (2004)
9. Margaritis, D., Thrun, S.: Bayesian network induction via local neighborhoods. In: Solla, S.A., Leen, T.K., Mller, K.-R. (eds.), Proceedings of the 1999 Conference 2000, vol. 12. MIT Press. pp 505-511 (2000)
10. Windeatt, T.: Ensemble MLP Classifier Design, vol. 137, pp.133–147. Springer, Heidelberg (2008)
11. Windeatt, T.: Accuracy/diversity and ensemble MLP classifier design. IEEE Transactions on Neural Networks 17(5), 1194–1211 (2006)
12. Yu, L., Liu, H.: Efficient feature selection via analysis of relevance and redundancy. J. Mach. Learn. Res. 5, 1205–1224 (2004)
13. 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)
14. Pudil, P., Novovicova, J., Kittler, J.: Floating Search Methods in Feature Selection. Pattern Recognition Letters, 15, 1119–1125 (1994)
15. Tsamardinos, I., Brown, L.E, Aliferis, C.F.: The max-min hill-climbing Bayesian network structure learning algorithm. Machine Learning, vol.65, pp 31–78 (2006)
16. Breiman, L.: Bagging predictors. Machine Learning, 24(2), 123–140 (1996)
17. Guyon, I.: Causality Workbench (2008) `http://www.causality.inf.ethz.ch/home.php`