Classification Using Markov Blanket for Feature Selection

Yifeng Zeng
Department of Computer Science
Aalborg University, Denmark
yfzeng@cs.aau.dk

Jian Luo
Department of Automation
Xiamen University, China
jianluo@xmu.edu.cn

Shuyuan Lin
Department of Computer Science
Fuzhou University, P.R.China
s030602108@fzu.edu.cn

Abstract

Selecting relevant features is in demand when a large data set is of interest in a classification task. It produces a tractable number of features that are sufficient and possibly improve the classification performance. This paper studies a statistical method of Markov blanket induction algorithm for filtering features and then applies a classifier using the Markov blanket predictors. The Markov blanket contains a minimal subset of relevant features that yields optimal classification performance. We experimentally demonstrate the improved performance of several classifiers using a Markov blanket induction as a feature selection method. In addition, we point out an important assumption behind the Markov blanket induction algorithm and show its effect on the classification performance.

1 Introduction

A classification task has found great challenges in the arising domains of biology, genetics and clinical diagnosis. The rapidly maturing technology offers a large data set for classification analysis that may contain thousands of features(or variables, we use features and variables interchangeably) with (in-)sufficient data instances. Unfortunately many of candidate features are either redundant or irrelevant to a target feature. Such data present familiar dimensional difficulties for classification of the target variable, and undermine the classification accuracy due to the noise of irrelevant variables. One solution is to preprocess the data set by selecting a minimal subset of variables and feed the selected features into a preferred classifier. It demands the work on implementing a feature selection method.

A good feature selection method picks an appropriate set of variables by pruning irrelevant ones from the data set. Much effort can be seen from a large amount of literature on various feature selection approaches. Details can be found in some comprehensive papers [1, 2, 3]. Two types of methods have been generally studied - *filter* methods and *wrapper* methods [4, 5]. A wrapper method iteratively evaluates the used classifier for every feature subset selected during the search while a filter method finds predictive subsets of features independently of the final classifier. Apparently, the wrapper method requires extremely expensive computation, which is infeasible in a large data set of interest. In this paper, we focus on the filter method.

The filter method using Markov blanket concept has received much attention in the current literature [6, 7]. The connection between the Markov blanket filter and other principled feature selection methods has been studied in [8]. The Markov blanket of a target variable contains a minimal set of variables on which all other variables are conditionally independent of the target variable. Most of the research work aims for a new Markov blanket discovering algorithm and then composes the final Markov blanket classifier [9, 10, 11].

In this paper, we do not aim for a new Markov blanket induction algorithm, but rather attempt to experimentally find out how the Markov blanket predictors could improve the classification accuracy of several well-studied classifiers like Naive Bayesian classifier [12], TAN [13], Id3 [14], C4.5 [15] and k-dependence Bayesian classifier(KDB) [16]. More importantly, we challenge the assumption behind the Markov blanket induction algorithm and study the effect on the classification accuracy. We select the Incremental Association Markov Blanket(IAMB) algorithm [10] for discovering a unique Markov blanket of the target variable. The IAMB algorithm is proved of correctness and exhaustively

experimented in learning Bayesian networks.

We organize this paper as follows. In section 2, we review the most relevant work on feature selection using Markov blanket concept. In section 3, we firstly present a generic classification framework that uses the Markov blanket in the preprocess stage for selecting relevant features. The IAMB algorithm is selected as the Markov blanket discovering algorithm. Finally, in section 4, we describe a comprehensive set of empirical results on demonstrating the improved performance of several classifiers given the Markov blanket predictors.

2 Related Work

Bayesian network (BN) [17, 18] is a directed acyclic graph where nodes represent variables of a subject of matter, and arcs between the nodes describe the causal relationship of variables. It is a compact representation of a joint probability distribution of domain variables. Markov blanket is a key concept of conditional independence in the graphical model of Bayesian networks. The Markov blanket of a target variable T is the set consisting of the parents of T, the children of T, and the variables sharing a child with T [17]. Given its Markov blanket, the variable T is conditionally independent from other variables in a BN. Formally, the Markov blanket of T is the minimal set of features conditioned on which all other features are independent of T. Similar to a large amount of work on learning Bayesian networks, research on Markov blanket discovery has been continuously updated since the first method of Grow-Shrink [9] appeared around ten years ago.

As implied by its name, the Grow-Shrink algorithm firstly finds the relevant variables to the target variable, and then reduces the estimated set of Markov blanket using conditional independence tests. However, the second stage of the Grow-Shrink algorithm is not proved sound. A similar line of work is the IAMB algorithm (We will describe it in length in Section 3.) that is proved correct and sound for discovering a unique Markov blanket [10, 19]. Later, many variants of the IAMB algorithm appeared to show better empirical results on the efficiency including Fast-IAMB [7], IAMBNPC [20], and parallel IAMB [21].

The value of Markov blanket concept has been early recognized in Koller and Sahami's work [6] on feature selection. Their work uses a heuristic method to identify the Markov blanket of a target variable. This indicator introduces more intensive research on designing Markov blanket classifiers and comparing their accuracy with other alternatives. One piece of such work presents the PCX classifier [22] relying on the PC algorithm [23] for identifying the Markov blanket gradually. Similar work investigates the variants of IAMB for composing classifiers and compares between them.

Our work contributes the growing line of work on Markov blanket classifiers. We study several classifiers using Markov blanket as feature selection method, and more importantly we point out the limitation of the enhanced classifiers that is challenged by the $faithfulness$ assumption behind the Markov blanket discovering algorithm.

3 The Framework

We describe a classification framework that accommodates both a generic classifier and a feature selection method using a Markov blanket induction algorithm. In this section, we detail the IAMB algorithm and discuss its limitations.

A general classification framework is illustrated in Fig. 1. Similar to the selection of classifiers, the feature selection method can use any of available Markov blanket induction algorithms. Note that the IAMB is a basic algorithm that is currently proved correct and sound, and is exhaustively tested in the literature [19].

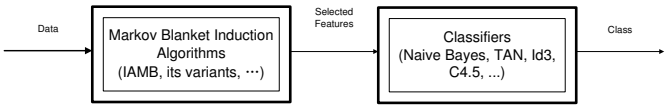


Figure 1. A framework classifies the input data instances by using a Markov blanket induction algorithm as feature selection. We select the IAMB algorithm in this paper. The intermediate results are selected features discovered by the IAMB, which is a minimal subset of relevant features.

The IAMB algorithm discovers a unique Markov blanket, $MB(T)$, of the target variable T (also called the $class$ variable in the classification) given data instances in the data set D. The algorithm takes an incremental strategy by starting an empty set and then gradually adding the Markov blanket elements. It comprises two steps: the growing phase and the shrinking phase. We show the IAMB algorithm in Fig. 2.

The growing step finds all possible nodes that have a strong dependency with the target variable T (lines 2-6). It measures the dependency using the conditional mutual information $MI(v, T|MB(T))$ (line 3). Other association functions could replace the information theoretical based measurement. To decide which variable shall be included into the initial Markov blanket,we use a conditional independence test (lines 4-5). One useful heuristic method finds initial Markov blanket variables through a threshold value [24]. The growing step terminates when no more new variables are added into the Markov blanket. It implies that

IAMB(T; D)

1: Assign $MB(T) = \emptyset$

Growing Phase: Add true positives to $MB(T)$

2: Repeat
3: Find v $max_{v \in (V - \{T\} - MB(T))} MI(v, T|MB(T))$
▷ V: a set of features in D
4: **IF** Not $CondInd(v, T|MB(T))$ **THEN**
5: $MB(T) = MB(T) \cup \{v\}$
6: Until $MB(T)$ does not change

Shrinking Phase: Remove false positives from $MB(T)$

7: For $v \in MB(T)$ **do**
8: **IF** $CondInd(v, T|MB(T) - \{v\})$ **THEN**
9: $MB(T) = MB(T) - \{v\}$
10: Return $MB(T)$

Figure 2. The IAMB algorithm utilizes two phases, $growing$ and $shrinking$, to find a correct Markov blanket for the target variable T.

the set is complete and no more variable could contribute the knowledge to the target variable T given the current Markov blanket.

As the computation of conditional mutual information relies on the set of Markov blanket that is formed as far (line 3), the false positives occurs in the growing phase. A variable that has the largest mutual information given the known Markov blanket is included in the current step; however, it would not be the one when the Markov blanket evolves over time. Thus, it is necessary to have the shrinking step to remove false positives from the built Markov blanket.

The shrinking step tests the conditional independence, $CondInd(v, T|MB(T) - \{v\})$, between each variable v and the target variable T given the remaining Markov blanket (line 8). Consequently, the false positives are removed from the Markov blanket gradually. We illustrate the two steps in one toy example below.

$Example$: Given the sampled data from Bayesian network (in Fig. 3), where c_5 is the target variable, we assume the order of node follows the list $\{c_1, c_4, c_2, c_3, c_6, c_7, c_8\}$. In the growing phase, we firstly add c_1 given the empty $MB(T)$. Subsequently, we add c_4 given $MB(T) = \{c_1\}$. Following similar procedures in the growing step, we include c_2, c_3, c_6, c_7 into the $MB(T)$. Finally, we don't count c_8 since it does not pass the conditional independence test.

The shrinking phase starts with the $MB(T) = \{c_1, c_4, c_2, c_3, c_6, c_7\}$. The first variable, c_1, is removed since it is conditional independent from T given $MB(T) = \{c_4, c_2, c_3, c_6, c_7\}$. Similar occurs to the variable c_4. Other

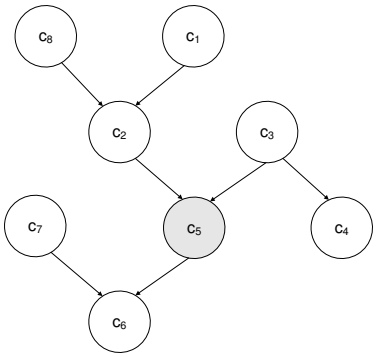


Figure 3. A toy example of Bayesian network illustrates the IAMB algorithm.

variables remain in the Markov blanket. The output is the reduced Markov blanket $MB(T) = \{c_2, c_3, c_6, c_7\}$. It is correct as shown in Fig. 3.

The correctness of the IAMB algorithm is proved in [10]. However, similar to other Markov blanket or Bayesian network discovery algorithms, the IAMB is sound under one important assumption $faithfulness$ [23]. A graph of Markov blanket or Bayesian network is faithful to a joint probability distribution over the set of features if and only if every dependence entailed by the graph is also present in the distribution (its details in [18]). In other words, if the distribution generated by the data is not $faithful$ to the graph, the IAMB algorithm may not ensure its optimality. We will show this limitation in the experiment.

The IAMB algorithm has the worst running time of $O(|V|^2 \times |D|)$ where $|V|$ is the number of features and $|D|$ the number of data instances. Many variants attempt to make the IAMB more efficient by chunking the set of features or ordering the test sequence in the growing phase [21].

4 Experimental Results

In this section, we report the results of analysis of the data from various sources. We firstly demonstrate the merit of the Markov blanket using the IAMB algorithm on the classification. We also compare between the correlation based attribute selection (CFs) [25] in Weka [1] by showing the performance averaged over cross-validation runs. Then, we use a couple of data sets to exhibit the limitations of using the IAMB in the feature selection oriented to the classification task.

Before proceeding to the result analysis, we describe three data sets and relevant parameter settings in Table 1.

[1] `http://www.cs.waikato.ac.nz/ml/weka/`

Classifier	All Features	IAMB	CFs
Naive Bayes	76.21%	**79.75%**	78.15%
TAN	78.32%	**79.98%**	78.91%
Id3	76.14%	**79.13%**	78.92%
C4.5	77.65%	79.29%	**79.76%**
KDB (K =2)	78.10%	**79.24%**	79.14%

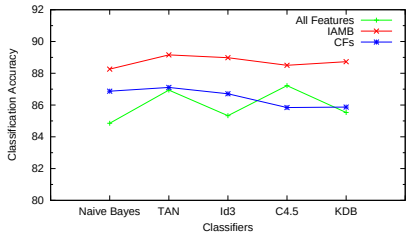
(a) Insurance Data Set

Classifier	All Features	IAMB	CFs
Naive Bayes	89.14%	91.12%	**91.23%**
TAN	90.21%	**93.31%**	88.47%
Id3	88.35%	**93.49%**	89.98%
C4.5	89.90%	**92.09%**	86.61%
KDB (K =2)	86.12%	**93.12%**	87.31%

(b) Alarm Data Set

Classifier	All Features	IAMB	CFs
Naive Bayes	89.20%	**93.90%**	91.22%
TAN	92.28%	**94.18%**	93.94%
Id3	91.49%	**94.32%**	91.23%
C4.5	**94.12%**	**94.12%**	91.14%
KDB (K =2)	92.38%	**93.84%**	91.17%

(c) Headache Data Set



(d) Performance Averaged over (a), (b) and (c)

Figure 4. (a, b, c) Performance of the framework either without pruning the features or using either the IMAB or CFs as the feature selection method. (d) summarizes the global performance by averaging results over the three data sets.

Dataset	Insurance	Alarm	Headache
Features	27	37	12
Sample Size	9000	6500	1000
Class	Accident	X35	Ha-Ho

Table 1. Data sets and the parameter settings in the IAMB improvement experiment.

The data sets are taken from benchmarks in the Bayesian research field. The insurance network estimates the expected claim cost for car insurance policyholder [26]. The alarm domain is compiled by a multiply connected Bayesian network representing the uncertain relationships among some relevant propositions in the intensive-case unit (ICU) [27]. The headache network [17] studies the involved causes yielding the corresponding headache symptom [2]. It is well known that causal relations exist among features within these three data sets. Hence, the $faithfulness$ assumption holds in executing the IAMB algorithm.

We experiment with the framework (in Fig. 1) on three data sets: $Insurance$, $Alarm$, and $Headache$. We also replace the IAMB with the CFs in the feature selection process that is noticed as one of effective attribute selection methods in Weka, and compare between them. In addition, we show the classification accuracy when no any attribute/feature is pruned from the data set. We use several general classifiers such as Naive Bayes, TAN, Id3, C4.5 and KDB [3]. The results are shown in Fig. 4

We observe that the classification accuracy is improved when the feature selection method is used to preprocess the data set. The performance of the IAMB outperforms the CFs selection method when both of them are used to prune the available features for the classification. To make a further investigation, we found that the CFs may not return a correct set of relevant features. For example, the CFs selected one more feature in the $Insurance$ data set. It may be due to the heuristic of the CFs that measures the relevance of features individually on the correlation to the class. We note that the classifiers differ in the classification accuracy [4]; however, their performance is improved steadily when the IAMB algorithm is used for selecting relevant features.

The correctness of the IAMB algorithm relies on the $faithfulness$ assumption. The assumption does not always hold in all types of data sets. Consequently, the IAMB algorithm could not produce correct features that are indeed

[2] `http://www.hugin.com`

[3] After several simulations, we set the parameter k =2 that gave the best results

[4] Discussion of classification performance of various classifiers is beyond the scope of this paper. Details in [28].

Classifier	All Features	IAMB	CFs
Naive Bayes	82.19%	86.18%	**87.98%**
TAN	81.23%	**86.92%**	86.18%
Id3	79.11%	81.76%	**83.90%**
C4.5	79.98%	**84.11%**	82.45%
KDB (K =2)	80.19%	82.75%	**85.91%**

(a) LRS Data Set

Classifier	All Features	IAMB	CFs
Naive Bayes	91.18%	92.34%	**92.93%**
TAN	91.98%	92.03%	**93.11%**
Id3	87.12%	86.11%	**87.23%**
C4.5	89.89%	88.38%	**89.97%**
KDB (K =2)	91.24%	92.78%	**93.47%**

(b) TIC Data Set

Figure 5. Performance of the framework when a data set does not obey the $faithfulness$ assumption.

Dataset	LRS	TIC
Features	85	86
Sample Size	531	5822
Class	FR49	Caravan

Table 2. Data sets and the parameter settings in the experiments in Fig. 5 .

relevant to the class or it may be insufficient to remove all irrelevant features, which definitely affects the final classification accuracy. We report such limitations over two data sets, LRS and TIC, in Fig. 5. Both the LRS and TIC data sets reside in the UCI repository [5]. The LRS data is a subset of the higher quality low resolution observations taken between 12h and 24h right ascension. The TIC data contains the complete set of possible board configurations at the end of tic-tac-toe games. Note that the domain knowledge implies no causal relations among features in the selected data sets.

The results show the IAMB algorithm is not so effective as the CFs method on improving the final classification performance. However, the benefit of using the IAMB is still visible since the selection method prunes many irrelevant or redundant features that introduce noise in the classification. To the best of our knowledge, this is the first piece of work empirically showing the limitations of the IAMB algorithm for the classification by challenging its assumption although some effort has been invested into a theoretical discussion.

5 Discussion and Conclusion

We have shown that the classification performance is improved when a Markov blanket induction algorithm is firstly used for pruning irrelevant features. It is mainly because the IAMB algorithm predicts a minimal set of truly relevant features. In contrast to previous work on Markov blanket classifiers, we do not directly use the discovered Markov blanket structure as a classifier for the final classification, but feed the Markov blanket predictors into a general classifier and empirically study its improvement. In addition, we challenge the $faithfulness$ assumption behind the Markov blanket induction algorithm. We shall note that most of studied domains may have causal relations between features so that the assumption is followed. However, we still need to be aware of possible consequences when a Markov blanket is misused as a feature selection method.

[5] `http://archive.ics.uci.edu/ml/`

Acknowledgement

The first author would like to thank Estrella Aparicio Garcia-Brazales for initiating this project when she visited Department of Computer Science in Aalborg University, Denmark.

References

[1] Langley, P.: Selection of relevant features in machine learning. In: In Proceedings of the AAAI Fall symposium on relevance, AAAI Press (1994) 140–144

[2] Dash, M., Liu, H.: Feature selection for classification. Intelligent Data Analysis **1** (1997) 131–156

[3] Liu, H., Yu, L.: Toward integrating feature selection algorithms for classification and clustering. IEEE Transactions on Knowledge and Data Engineering **17** (2005) 491–502

[4] Almuallim, H., Dietterich, T.G.: Learning with many irrelevant features. In: In Proceedings of the Ninth National Conference on Artificial Intelligence, AAAI Press (1991) 547–552

[5] Kohavi, R., John, G.H.: Wrappers for feature subset selection. Artificial Intelligence **97**(1-2) (1997) 273–324

[6] Koller, D., Sahami, M.: Toward optimal feature selection. In: Proceedings of the Thirteenth International Conference on Machine Learning, Morgan Kaufmann (1996) 284–292

[7] Yaramakala, S., Margaritis, D.: Speculative markov blanket discovery for optimal feature selection. In: ICDM '05: Proceedings of the Fifth IEEE International Conference on Data Mining, Washington, DC, USA, IEEE Computer Society (2005) 809–812

[8] Tsamardinos, I., Aliferis, C.F.: Towards principled feature selection: Relevancy, filters and wrappers. In: in Proceedings of the Ninth International Workshop on Artificial Intelligence and Statistics, Morgan Kaufmann Publishers (2003)

[9] Margaritis, D., Thrun, S.: Bayesian network induction via local neighborhoods. In: in Advances in Neural Information Processing Systems 12 (NIPS, MIT Press (1999) 505–511

[10] Tsamardinos, I., Aliferis, C.F., Statnikov, E.: Algorithms for large scale markov blanket discovery. In: In The 16th International FLAIRS Conference, St, AAAI Press (2003) 376–380

[11] Aliferis, C.F., Tsamardinos, I., Statnikov, E.: Hiton, a novel markov blanket algorithm for optimal variable selection. In: The 2003 American Medical Informatics Association (AMIA) Annual Symposium. (2003) 21–25

[12] Ripley, B.D.: Pattern Recognition and Neural Networks. Cambridge University Press, U.K. (1996)

[13] Friedman, N., Goldszmidt, M.: Building classifiers using bayesian networks. In: In Proceedings of the thirteenth national conference on artificial intelligence, AAAI Press (1996) 1277–1284

[14] Quinlan, J.R.: Induction of decision trees. Machine Learning **1**(1) (1986) 81–106

[15] Quinlan, J.R.: C4.5: programs for machine learning. Morgan Kaufmann (1993)

[16] Sahami, M.: Learning limited dependence bayesian classifiers. In: In KDD-96: Proceedings of the Second International Conference on Knowledge Discovery and Data Mining, AAAI Press (1996) 335–338

[17] Jensen, F.V.: An introduction to Bayesian networks. Springer, New Work (1996)

[18] Pearl, J.: Probabilistic reasoning in intelligent systems: networks of plausible inference. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA (1988)

[19] Tsamardinos, I., Brown, L.E., Aliferis, C.F.: The max-min hill-climbing bayesian network structure learning algorithm. Machine Learning **65**(1) (2006) 31–78

[20] Aliferis, C., Tsamardinos, I.: Algorithms for large-scale local causal discovery and feature selection in the presence of small sample or large causal neighborhoods. In: Technical Report DSL 02-08, Department of Biomedical Informatics, Vanderbilt University, Nashville, TN, USA (2002)

[21] Aliferis, C., Tsamardinos, I., Statnikov, A.: Large-scale feature selection using markov blanket induction for the prediction of protein-drug binding. In: Technical Report DSL TR-02-06, Department of Biomedical Informatics, Vanderbilt University, Nashville, TN, USA (2002)

[22] Bai, X., Glymour, C., Padman, R., Ramsey, J., Spirtes, P., Wimberly, F.: Pcx: Markov blanket classification for large data sets with few cases. In: CMU-CALD-04-102, School of Computer Science, Carnegie Mellon University (2004)

[23] Spirtes, P., Glymour, C., Scheines, R.: Causation, Prediction, and Search. Springer, New Work, USA (1993)

[24] Garcia-Brazales, E.A.: Discovering the Markov blanket as feature selection oriented to classification task. Master Thesis, Department of Computer Science, Aalborg University, Denmark (2007)

[25] Hall, M.A.: Correlation-based Feature Subset Selection for Machine Learning. PhD Thesis, University of Waikato, Hamilton, New Zealand (1998)

[26] Binder, J., Russell, S., Smyth, P.: Adaptive probabilistic networks with hidden variables. **29**(2) (1997) 213–244

[27] Beinlich, I.A., Suermondt, H.J., Chavez, R.M., Cooper, G.F.: The alarm monitoring system: A case study with two probabilistic inference techniques for belief networks. In: European Conference on Artificial Intelligence in Medicine. (1989) 247–256

[28] Geiger, D., Goldszmidt, M., Provan, G.: Bayesian network classifiers. In: Machine Learning. (1997) 131–163