ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 6, June 2014

A Comparison of Filter and Wrapper Approaches with Data Mining Techniques for Categorical Variables Selection

Bangsuk Jantawan[1, 2], Cheng-Fa Tsai[2]

Department of Tropical Agriculture and International Cooperation, National Pingtung University of Science and Technology, Pingtung, Taiwan[1]

Department of Management Information Systems, National Pingtung University of Science and Technology Pingtung, Taiwan[2]

ABSTRACT: The purpose of this study is to evaluate the most important features of graduate employability in higher education database, in attempt to measure the employability situation for graduate information of the Maejo University in Thailand. The experiment also applies the features selection methods to increases the overall efficiency of classification model. There are two general attribute selection approaches: the Filter approach and the Wrapper approach. The Filter approach includes 3 methods, including Information Gain, Gain Ratio and Chi-square. The Wrapper approach we used Search method consisting of Genetic Search, Best First search and Greedy Stepwise as random search approach for subset generation, wrapped with different bayesian classifiers namely Naïve bayes, Bayes network with K2 algorithm, Bayes network with TAN algorithm and Bayes network with Hill-climber algorithm. The results illustrate, employing feature subset selection using proposed wrapper approach has enhanced classification accuracy.

KEYWORDS: Attribute Selection, Data Mining, Filter Approach, Wrapper Approach

I. INTRODUCTION

Nowadays, amount of data has rapidly increased every year, especially in term of graduate employability historical database. It contains huge amounts of data in database. These data are also including currently unknown and potentially interesting patterns and relations, which can be uncovered using knowledge discovery and data mining techniques. Data mining techniques have enormous potential for analysing the hidden patterns in the data sets of the education domain. These patterns can be utilized for the decision making of administrator in higher education. Data pre-processing is an essential phase in the knowledge discovery process, once quality decisions must be based on quality data [1]. This process, when applied early to mining, can substantially improve the overall quality of the patterns mined and/or the time required for the actual mining [1].

The aim of data reduction is to search a minimum set of features such that the resulting likelihood distribution of the output (class) is as close as possible to the original distribution obtained using all features. Mining on the reduced set of features has increasing benefits. It reduces the number of features appearing in the discovered patterns, in order to help and make the patterns easier to understand. Then, it enhances the classification preciseness and learning runtime. Section 2 explains the related work on feature selection method, filter approach, and wrapper approach. Methodology and method are described in section 3 followed by results and conclusion in section 4 and 5 respectively.

II. RELATED WORK

Feature selection is the processes that choose a subset of relevant features for building the model. Feature selection is one of the most important and frequently used techniques in data preprocessing for data mining [1]. It is also useful in term of the data analysis process, as it shows which the input variables or features are important for predicting, and how those features are related. The goal of feature selection for classification task is to maximize classification accuracy [2]. It is easier to choose features for classification than for clustering, once the classification uses class label information. Although, domain experts can eliminate few of the irrelevant attributes, choosing the best subset of features usually requires a systematic approach [1].

Feature selection can be divided in filter methods and wrapper methods. Filter methods is defined as using some actual property of the data in order to select feature using the classification algorithm. Entropy measure has been used as filter method for feature selection for classification [3]. Wrapper approaches apply classification algorithm to each candidate feature subset and then evaluate the feature subset by threshold functions that utilize the classification result. Moreover, wrapper method can combine a classification technique with a Bayesian inference mechanism for automatically selecting relevant features. Feature selection methods provide three main benefits when building predictive models as following: (1) the model improving is obviously interpretation; (2) they can make shorter training times; and (3) enhanced generalization by reducing over fitting.

A. *Filter Method*

The filter method precedes the actual classification process. The filter approach is input variables of the learning induction algorithm, computationally simple and rapidly scalable. Using filter method, feature selection is done once later can be provided as input to different classifiers [1]. Various feature selection techniques and feature ranking have been offered such as Correlation-based Feature Selection (CFS), Gain Ratio (GR), Chi-Squared, and Information gain etc.

Chi-Squared [5] is based on the χ2-statistic and evaluates each feature input variables with respect to the class labels (output). The larger the Chi-squared, the more relevant the feature is with respect to the class. Obtained the number of intervals (I), the number of classes (B), and the total number of instances (N), the Chi-squared value of a feature is computed as following:

$$\chi 2 = \sum_{i=1}^{I} \sum_{j=1}^{B} \frac{\left[A_{ij} - \frac{R_i * B_j}{N} \right]^2}{\frac{R_i * B_j}{N}} \qquad (1)$$

where R_i means the number of instances in the i^{th} interval;
B_j is the number of instances in the j^{th} class; and
A_{ij} is the number of instances in the i^{th} interval and j^{th} class.

Information Gain [5], Information Gain gauges the number of bits of information gained about the class prediction when using a given feature to assist that prediction [5]. For each feature, a score is obtained based on how much more information about the class is gained when using that feature. The information gain of feature X is defined as following:

$$Information\ Gain\ (X) = H(Y\ ) - H(Y\ |X) \qquad (2)$$

where $H(Y\ )$ and $H(Y\ |X)$ mean the entropy of Y and the conditional entropy of Y given X, respectively.

The level of a feature essential is thus determined by how great is the decrease in entropy of the class when considered with the relevant feature individually.

Gain Ratio [5] is a refinement to Information Gain. While Information Gain favors features that have a large number of values, The Gain Ration approach is to maximize the feature information gain while minimizing the number of its values. The gain ratio of X is thus defined as the information gain of X divided by its intrinsic value:

$$Gain\ Ratio\ (X) = IG(X)/IV\ (X) \qquad (3)$$

where

$$IV\ (X) = \sum_{i=1}^{r} \left(\frac{|Xi|}{N}\right) \log\left(\frac{|Xi|}{N}\right)$$

from which $|Xi|$ is the number of instances where attribute X takes the value of Xi;
r is the number of distinct values of X; and
N is the total number of instances in the dataset.

B. *Wrapper Approach*

The wrapper approach, Isabelle Guyon and Andre Elisseeff [4] offers a simple and effective way to problem solution of variable selection, regardless of the selected learning machine. The method of the learning machine lends to use off-the –shelf machine learning software packages. In the general formulation, the wrapper approach consists in using the prediction performance of a given learning machine to evaluate the relative usefulness of subsets of variables.

Moreover, in practice, one needs to define: (1) How to find the space of all possible variable subsets; (2) How to evaluate the prediction performance of a learning machine to guide the search and halt it; and (3) which predictor to use. An exhaustive search can probable be performed, if the number of variables is not large too much. In addition, wide range of search strategies can be used, including branch-and-bound, best-first, genetic algorithms simulated annealing (see [5], for a reviews). Performance evaluation are usually done using a validation set or by cross-validation. As shown in this special issue, widespread predictors include naïve Bayes, decision tree, and support vector machines.

III. MATERIALS AND METHODS

A. *Datasets and Data Pre-processing*

1) Dataset

The case university considered in this paper is the Major University in Thailand. The dataset used in this experiment was obtained data within the graduate employability database in three academic years of the Planning Division Office, Maejo University. The first academic year is 2009, graduate employability from three degrees was collected: the total number of bachelor degree is 3,356, master degree 156 and doctoral 17 persons. In the academic year of 2010: the total number of bachelor degree is 3,947, master degree 119 and doctoral 20. In the last academic year is 2010: the total number of bachelor degree is 4,156, master degree 83 and doctoral 3 graduates.

2) Data Pre-processing

- Year (YEAR): 2009 (Y2009), 2010 (Y2010), 2011 (Y2011)
- Gender (GEND): Male (M), Female (F)
- Domicile (Domicile): 76 provinces in Thailand such as Suratthani province (Suratthani), Chiang Mai province (ChingMai) and etc.
- Degree (DEG): Doctorate (Doctoral), Master degree (Master), Bachelor degree (Bachelor)
- Work province (WorkPro): 76 provinces in Thailand such as Suratthani province (Suratthani), Chiang Mai province (ChingMai) and etc.
- Educational background (ED): Bachelor of Science (BSc), Bachelor of Landscape Architecture (BLA), Bachelor of Engineering (BEng), BA Economics (BEcon), Bachelor of Business Administration (BBA), Doctor of Philosophy (PhD), Bachelor of Arts (BA),Bachelor of Agricultural Technology (BSPlantScience), Bachelor of Accountancy (BAcc), Bachelor of Political Science (BPolSc), Master of Science (MSc), Master of Business Administration (MBA), Master of Engineering (MEng), Doctor of Arts (PhDArts), Master of Arts (MArts), Bachelor of Technology (BTech)
- Faculty (ISCED): Faculty of Agricultural Production (FaOAgPr), Administration (FaOAd), Faculty of Science (FaOSc), Faculty of Engineering and Agro-Industrial (FEnA), High School- Phrae Honor (HSPhr),High School – Chumphon (HScChu), College of Management Sciences (CoMaSc), The School of

Renewable Energy (ScReEn), Faculty of Tourism Development (FOToDe), Faculty of Liberal Arts (FOLiBr), Faculty of Economics (FaOEc), Faculty of Fisheries an Aquatic Resources (FOFiAq), Faculty of Information and Communication (FaOInACo), Faculty of Architecture and Environmental Design (FaOArAEnDe), Faculty of Animal Science and Technology (FaOAnScATe)

- Grade Point Average (GPA): Numerical data (1.00, 2.50, …)
- Talent (TAL): Foreign languages (ForeL), Computer (Comp), The recreational activities (TReAc), Arts (ART), Sports (Spor), The Performing Arts / Music chorus (PAMC), Have Not (NO), Other (OT)
- Position (JOB): Officials / authorities of the state/State enterprise (OASE), Companies / organizations/ private businesses(COPB), Independent Business / Owner (IBO), Employees bodies / international (EBI), Other (OT), Not specified(No)
- Length of time for finding job (TFJ): Unknown (Nu), Get a job immediately after graduation (IMME), 1-3 months (M13), 4-6 months (M46), 7-9 months (M79), 10-12 months(M1012), Over 1 year (Over1Y), Work before and during the study (BNS)
- The matching between field of graduate and job (MAF): Match (Yes), Not Match (No), Unknown(Un)
- Required for studying (ReSt): Demand (Yes), Needless (No)

Included also was a response factor called Status (STA) which could take an Unemployed (UnE), Unemployed and Study (UnES), Employed and Study (ES), Employed (E) value.

B. *Methods*

This part was separated into two parts as follows: the first part in which the different methods were used in order to select highly relevant features, and a second part that consisted of applying – using the highly related features – different data mining techniques in order to essay their accuracy for predicting graduate status. The tools in this experiment used the Waikato Environment for Knowledge Analysis (WEKA) version 3.6.10 freeware which is a popular suite of machine learning software written in Java, developed by the University of Waikato, New Zealand.

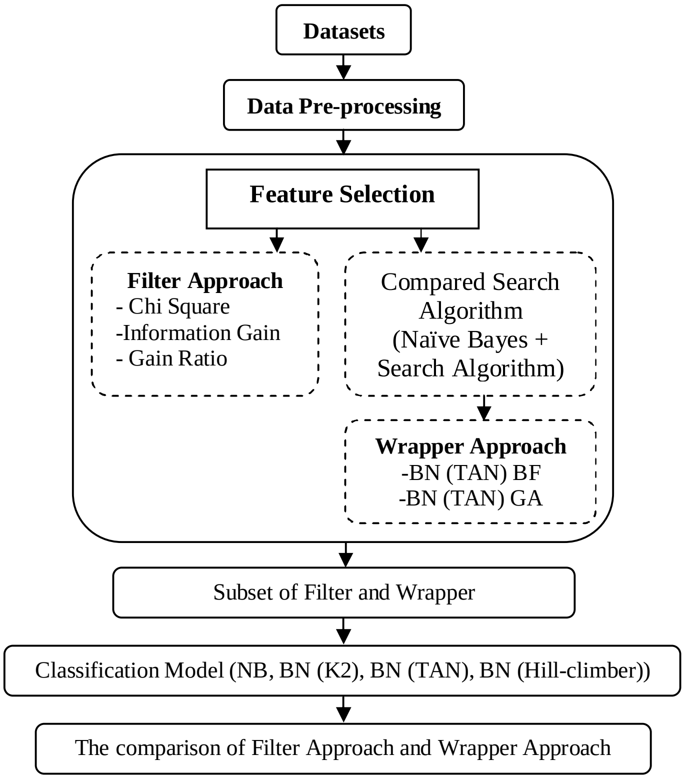
Datasets
Data Pre-processing
Feature Selection
Filter Approach
- Chi Square
-Information Gain
- Gain Ratio
Compared Search Algorithm (Naïve Bayes + Search Algorithm)
Wrapper Approach
-BN (TAN) BF
-BN (TAN) GA
Subset of Filter and Wrapper
Classification Model (NB, BN (K2), BN (TAN), BN (Hill-climber))
The comparison of Filter Approach and Wrapper Approach

Fig 1. Steps for feature selection

1) Feature Selection

This phase was offered in order to identify highly feature related to 13 attribute in term of predicting the class label. WEKA software provides a number of attribute selection methods for studying the relevant of database factors in related with a feature response.

Our experiment using cross-validation was conducted with ten-fold cross-validation under the WEKA attribute selection function. The function (select attributes) grouped different methods in two distinct categories based on either which shows the detail below.

Table 1. Feature evaluation methods, Classifier and search methods under WEKA data mining software for feature selection

Feature evaluation methods	Search Method	Classifier/options
Wrapper method		
WrapperSubSetEval	Greedy	BayesNet (BN), algorithm K2,TAN
	BestFirst	BayesNet (BN), algorithm Hill-climber
	Genetic	Naïve Bayes (NB), Pre-defined values
Filter methods		
ChiSquareAttributeEval	Ranker	Pre-defined set
GainRatioAttributeEval	Ranker	Pre-defined set
InfoGainAttributeEval	Ranker	Pre-defined set

Wrapper method: The frequency with which each feature was selected in 10 folds cross-validation training, which we used. Wrapper method, which is using a classifier plus cross-validation for evaluate attribute sets by using a learning (a classifier plus cross-validation) scheme. That is selecting attribute on the basis of improvement in predictive ability brought about by the incorporation of each feature.

Filter method: An indicator of the merit of each attribute that provided a ranking of factors in terms of correlation. There are four such selection models were used.

We applied different data mining models after the attributes had ranked according to weight-relevance for the graduate status prediction factor so as to ensure that in a percentage of correctly classified cases and a confusion matrix for each.

We also used three different ways for testing the model to corroborate that the number of variables had a bearing on model predictive abilities. The data mining techniques were used to construct the model of predicting attributes on the variable response (graduate status) which is Bayesian networks.

Bayesian networks: Bayesian networks are directed acyclic graphs used for descriptive and predictive purposed. Their node-and-arc network structure provides information on independence/dependence relationships (depicted by arcs) and factors (depicted by nodes). For our research we used K2, hill-climber and TAN implemented in WEKA as the network training algorithms, with different constraints on the number of parents. We also used a particular case of the Baysesian networks, called naïve Bayes, with a structure of just two levels and a single parent (the response variable) pointing to all the co-variables. The networks were trained by means of a greedy search of the space of possible structures, with the best network chosen on the basis of a specific goodness-of-fit criterion for the selected algorithm [6].

K2 algorithm: This algorithm, which uses a greedy search mechanism, starts from the simplest, possible network, which each successive repetition, is modified by the addition of new parents producing greater benefits according to a pre-established criterion [6].

Hill-climber algorithm: this algorithm, which is a discrete version of the gradient descent (ascent) algorithm, implements a local search of each repetition. The algorithm starts with an initial network and determines a nearest-neighbor graph that improves the network by including, eliminating or inverting an arc in the graph. The process is repeated until there is no neighbor that improves the current solution [7].

TAN algorithm; this algorithm first constructs the attributes tree structure and then adds the class variable following the nave structure [8].

IV. EXPERIMENTAL RESULTS

Results for every stages of our experiment performed on the graduate historical data in higher education database are presented below.

The different feature evaluation methods provided relatively similar results. Filter approach with Chi-Square, Information gain and Gain ratio were applied together with naïve Bayes, Bayesian with algorithm K2, Bayesian with algorithm Hill-climber, and Bayesian with algorithm Tan. The result shows in Table 2.

Table 2. Classification using filter feature selection approach

Filter	Feature	NB	BN(K2)	BN(Hill-climber)	BN(TAN)
ChiSquar	13	94.11	94.22	94.02	**94.56**
InfoGai	13	94.11	94.22	94.02	**94.56**
GainRat	13	94.11	94.22	94.02	**94.56**

Wrapper approach with Best First, Genetic and Greedy Stepwise were used as random search approaches wrapped with different Bayesian network classifiers namely naïve Bayes, Bayesian with algorithm K2, Bayesian with algorithm Hill-climber, and Bayesian with algorithm Tan. The result shows in Table 3.

Table 3. Classification using search method selection approach

Search Method	Feature	NB	BN(K2)	BN(Hill-climber)	BN(TAN)
Best First	9	94.64	94.66	94.71	**95.10**
Genetic Search	9	94.64	94.66	94.71	**95.10**
Greedy Stepwise	1	94.56	94.56	94.56	94.56

Table 4 shows the reduced relevant attributes identified by different wrappers: Genetic search + Bayesian with algorithm Tan, and Best first search + Bayesian with algorithm Tan for improving classification accuracy of different classifiers in validation step. Validation was done using four classifiers namely naïve Bayes, Bayesian with algorithm K2, Bayesian with algorithm Hill-climber, and Bayesian with algorithm Tan.

Table 4. Classification accuracy using Bayes Network (Tan algorithm) together with Genetic and Best First search

Method	Feature	Classifiers Accuracy (%)			
		NB	BN(K2)	BN(Hill-climber)	BN(TAN)
GEBN(TAN)	9	94.64	94.66	94.71	**95.10**
BFBN(TAN)	9	94.64	94.66	94.71	**95.10**

V. CONCLUSIONS

As for the results have enabled us, using different data mining approach under Bayesian network, in order to determine the efficiency scheme for spending and analyzing graduate employability dataset that identifies the most immediately correspond attributes and resulting against prediction success rates or accuracy rate. The scheme is executed in two parts. First of all, the most important attributes are pre-selected using various methods. In our experiment, all the methods used produced similar accuracy results. The results of this paper represent an important advance of managing data on graduate employability. The satisfactory results of Bayesian networks with the Tan algorithm indicate these to be reliable tools for studies of graduate employability and their accuracy.

ACKNOWLEDGMENTS

B. Jantawan would like to express thanks Dr. Cheng-Fa Tsai, professor of the Management Information Systems Department, the Department of Tropical Agriculture and International Cooperation, National Pingtung University of Science and Technology in Taiwan for supporting the outstanding scholarship, and highly appreciates to Mr. Nara Phongphanich and the Planning Division Office, Maejo University in Thailand for giving the information.

REFERENCES

1. Asha Gowda Karegowda, M.A.Jayaram and A.S. Manjunath, "Feature Subset Selection Problem using Wrapper Approach in Supervised Learning", International Journal of Computer Applications, Vol. 1, No. 7, pp. 0975–8887, 2010.
2. Ron Kohavi, George H. John, "Wrappers for feature subset Selection", Artificial Intelligence, Vol. 97, No. 1-2. pp. 273-324, 1997.
3. Dash, M. and H. Liu, "Feature Selection for Classification", Intelligent Data Analysis, Vol. 1, pp. 131–156, 1997.
4. Isabelle Guyon and Andre Elisseeff, "An Introduction to Variable and Feature Selection", Journal of Machine Learning Research, Vol. 3, pp. 1157-1182, 2003.
5. Cooper, G.F. and Herskovits, E.A., "Bayesian method for the induction of probabilistic networks from data", Machine Learning, Vol. 9, No. 4, pp. 309-47, 1992.
6. Nocedal, J. and Wright, S.J., "Numerical optimization", Springer, 1999.
7. Friedman, N., Geiger D. and Goldszmidt, M., "Bayesian network classifiers", Machine Learning, Vol. 29, No. 2-3, pp. 131-63, 1997.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 6, June 2014

BIOGRAPHY

Bangsuk Jantawan is a Ph.D. student in the Department of Management Information System at the National Pingtung University of Science and Technology, Taiwan. Her research is the Application of Data Mining to Build Classification Model for Predicting Graduate Employment. Ms. Jantawan present research interests include the data mining, education system development, decision making, and machine learning.

Cheng-Fa Tsai is a full professor in the Department of Management Information Systems (MIS), National Pingtung University of Science and Technology, Taiwan. He has published over 160 well-known journal papers and conference papers and several books in the MIS. He holds, or has applied for, nine U.S. patents and thirty ROC patents in his research areas. Prof. Dr. Tsai research interests are in the areas of education, data mining and knowledge management, database systems, mobile communication and intelligent systems, with emphasis on efficient data analysis and rapid prototyping.