Identifying Fewer Key Factors by Attribute Selection Methodologies to Understand the Hospital Admission Prediction Pattern with Ant Miner and C4.5

Kyoko Fukuda

Geo Health Lab, Department of Geography, University of Canterbury, Private Bag 4800, Christchurch, New Zealand
kyoko.fukuda@canterbury.ac.nz

Abstract. Attribute Selection (AS) is generally applied as a data pre-processing step to sufficiently reduce the number of attributes in a dataset. This study uses six different data mining AS methods to identify a few key driving climate and air pollution attributes from small attribute sets (16 attributes) to increase knowledge about the underlying structures of acute respiratory hospital admission counts, because understanding key factors in environmental science data helps constructing a cost effective data collection and management process by focusing on collecting and investigating more representative and important variables. The performance of the selected attribute set was tested with Ant-Miner and C4.5 classifiers to examine the ability to prediction the admission count. Removal of attributes was successful over all AS methods, especially TNSU (a newly developed AS method, Tree Node Selection for unpruned), which achieved best in removing attributes and some improving the classification accuracy for Ant-Miner and C4.5. However, the overall prediction accuracy improvements are small, suggesting that AS selects attribute sets sufficiently enough to maintain the accuracy for Ant-Miner and C4.5.

Keywords: Attribute Selection, Ant Miner, Air pollution, Hospital admission.

1 Introduction

Attribute Selection (AS) methods are generally applied to practically reduce computation time on large data sets, e.g., the thousands of attributes that are often seen in text or web mining problems, whereas attribute sets in environmental science can be reasonably small, e.g., from ten to less than hundreds, due to availability and accessibility, e.g., historically unmeasured variables or limitations in setting up monitoring sites. AS acts as a knowledge discovery tool by ranking the importance of attributes (ranking filter method) or selecting a subset of attributes (subset evaluator) [1, 2, 3]. This study uses four commonly known AS methods in WEKA [4]; Information Gain Attribute Ranking (IG) [5], Relief (RLF), e.g., [6], Correlation-based Feature Selection (CFS) [2], Consistency-based Subset Evaluation (CNS) [7] and a newer AS method; Tree Node Selection (TNS) for unpruned (TNSU) and for pruned (TNSP) [1], to identify fewer but key climate and air pollution factors to understand the underlying structures of acute respiratory hospital admission patterns.

J.D. Velásquez et al. (Eds.): KES 2009, Part II, LNAI 5712, pp. 853–860, 2009.

Performance of each selected attribute set (input) was then tested with the Ant-Miner [8] and C4.5 [5] algorithms to observe prediction ability of admission pattern (class; *yes* for two or more admission counts otherwise *no*, which follows *outbreak* definition [9]). While the acute respiratory admissions are commonly known to be affected by changes in various air pollution and climate levels and generally investigated by statistical analysis [10], the goal of this study is to determine key climate and air pollution factors for respiratory hospital admissions from applying AS methods rather than reducing the attribute dimensions, which is a new approach in studying air pollution and health. To the best of our knowledge, these data mining attribute selection methodologies have not been commonly examined as data pre-processing methodologies or for selection of key attributes to obtain improved results for the Ant-Miner classifier. In particular, solving environmental science problems by using a combination of Ant-Miner and attribute selection is not yet a common procedure.

The next method section briefly outlines each AS method, followed by previous AS in Ant-Miner, the study data, in addition to introducing the motivation for producing the admission prediction model for the study site, and how AS methods would help. The final two sections present results and discussions, and summarize the finding as conclusions. The aim of this paper is to introduce applications of various AS methods with Ant-Miner and C4.5 algorithms for a real environmental science case study. The brief descriptions of AS, Ant-Miner and C4.5 algorithms are discussed or omitted as details are shown in [1, 2, 4, 5, 8, 11].

2 Methods

2.1 Background of Attribute Selection Methods

Detailed performance of AS methods, e.g., classification accuracy and reduction of attributes, that were used in this study were previously investigated by [1] for testing IG, RLF, CFS, CNS and wrapper methods to propose TNSP and TNSU, and [2] for testing IG, RLF, CNS, principal component analysis and wrapper methods to propose CFS, from applying naïve Bayes and C4.5 algorithms on various sized benchmarking databases. For example, reference [1] tested 33 benchmark datasets from 4 attributes to less than 300 attributes and from 40 to less than 50,000 instances. The selected attribute set varies depends on how each algorithm works; see details in [1, 2, 4]. The simplest attribute selection method [2] is IG [5], which quickly selects and orders attributes by importance by measuring the information gain in respect to the class, RLF [6] searches important attributes by repeatedly selecting a randomly selected instance from its two nearest neighbours between the same class and others, CNS [7] looks for a subset of attributes with the best consistency in class values, CFS [2] searches for a good subset of attributes by considering the usefulness of individual features at predicting each class, and a newer attribute selection method, Tree Node Selection (TNS), identifies a set of attributes using a pre-constructed decision tree as an information source by counting numbers of instances that go through nodes and classes. For example, the root node of a decision tree tends to be most important as it connects to the rest of the nodes to distribute instances to classes, but if the attribute

was more frequently used to construct many leaf nodes, then that attribute can also be important; see details in [1].

From the benchmark experiments, references [1] and [2] found that the wrapper is the best AS method, but it is time consuming. Reference [1] found that TNS performed consistently in reduction of attributes and obtained high accuracy over various data mining attribute selection approaches, whereas other methods tend to trade off performance in reduction of attributes and accuracy. Reference [2] found that CFS, CNS and RLS are good overall, but there is a trade off in performance among them.

2.2 Attribute Selection with Ant-Miner

The Ant Colony Optimization (ACO) algorithm is a swarm intelligence technique that mimics real ant behaviour. Recently, ACO has been used to solve attribute or feature selection problems. Reference [12] used ACO in rough set theory [13] to obtain high accuracy and minimum sets of features, since ACO finds solutions rapidly with very small cardinality during its pheromone update rule and solution construction process. Reference [14] developed an ACO feature selection method and tested it on a text categorization problem against other data mining and statistical attribute selection methodologies, Information Gain (IG), χ^2 statistics (CHI) and genetic algorithms (GA), by performing nearest neighbour classification. They found that ACO outperformed IG and CHI, and GA is almost comparable to ACO in terms of maintaining the accuracy and selecting minimum feature subsets. They stated that for datasets with more features, ACO has a strong search capability in the problem space, as a search continues until the optimal solution is found, whereas GA cannot find a better one after finding a sub-optimal solution.

2.3 Ant-Miner and C4.5 Classifiers

In order to test performance of attribute sets selected by AS methods, Ant-Miner and C4.5 algorithms were used to compare the prediction ability of admission counts using the smaller sets of selected attributes. Since this study is a preliminary investigation, the traditional Ant-Miner [8] has been used, because it is still a flexible and robust classification mining method which works well [15, 16], even though newer Ant-Miner algorithms have been developed, e.g., Ant-Miner 2 [17], ACO-Miner [11] and TACO-Miner [16]. In comparison to Ant-Miner, one of the most well known classification algorithms, C4.5 [5], was tested because Ant-Miner [8] is similar to a decision tree algorithm that discovers classification rules by following a divide-and-conquer approach:

IF < term1 and term2 and ...> THEN <class>

However, the heuristic functions for decision tree algorithms and Ant-Miner differ in how they consider the entropy; for the former they are computed for an attribute as a whole, but the latter computes them for an attribute-value pair only [8]. For Ant-Miner and C4.5 used the default parameter setting of Ant-Miner software [8] and WEKA [4] respectively was used.

2.4 Attribute Selection Process

Fig. 1 describes attribute selection steps. Firstly, the entire data (full attribute set) was divided into 90% (from the start of the studied period) and 10% (towards the end of studied period) to create the training and test set. Ant-Miner and C4.5 classifiers are applied on the training set to obtain the *original* classification accuracy (before the AS process) via the 10-fold cross validation process. Secondly, two sets of AS approaches are carried out. Ranking filter approaches (TNSP, TNSU, IG and RLF) ranks each attribute by its importance, i.e., the top labelled rank from "1" indicates the most important attribute and so on. Each set of ranked attributes separately runs the Ant-Miner and C4.5 classifiers, iteratively removing the least important attribute one by one to obtain the classification accuracy and the process continues until a single attribute remains. Subset evaluator approaches (CFS and CNS) select the attribute set at once, whereas the attribute set that obtained the highest accuracy for the ranking filter is used for the prediction process. Ant-Miner and C4.5 are separately run with the selected subset of attributes from the training set to extract classification rules. The created rules are then tested on the test set (unknown data points, not used to select the attribute set) to obtain the prediction accuracy, the *final* classification accuracy, of the respiratory admission pattern.

To identify key climate and air pollution factors for the admission pattern, the obtained attribute set (with the highest classification accuracy) is examined and compared among AS methods. Here, the top 3 commonly selected attributes throughout all AS methods will be summarised as follows. When TNSU ranks TG at rank 1, it gives one point to TG. If another AS method ranked TG at rank 2, it adds another point to TG. The total points are added up. The attribute that records the highest score is considered to be the most frequently selected attribute over all AS methods. Note that all points are counted equally as "one point" regardless of rank, i.e., first or third rank.

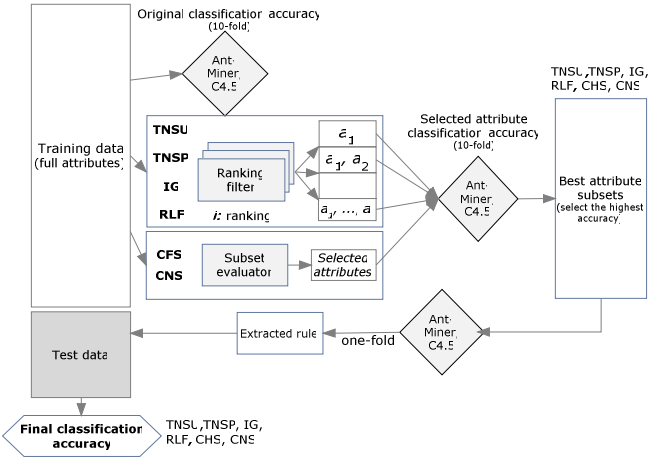

Fig. 1. Attribute selection processes for Ant-Miner and C4.5 classifiers

Relative proportion of selected number of attributes (in %) is calculated to compare the attribute reduction performance among all AS methods. The two sample means (and standard deviation) of the *original* and *final* classification accuracy of Ant-Miner and C4.5 are calculated to assess how the means of classification accuracy differ among classification algorithms.

2.5 Studied Data

The study area, Christchurch, New Zealand, suffers from severe air pollution problems in winter due to domestic heating by burning wood, e.g., [10]. The studied area, Christchurch City, is located in the South Island of New Zealand. The main winter air pollutants in Christchurch are carbon dioxide (CO_2) from domestic heating and motor vehicles, particulate matter (PM and PM_{10}, particles of diameter 10 micrometers or less) from domestic heating, sulfur dioxide (SO_2) from industry, e.g., [10]. Some pollutants can record beyond the acceptable air pollution guideline during winter due to the heavy use of wood for domestic heating. It is desirable to promote a good prediction method for the outbreak acute admission rate in order to help with the hospital care management.

A total of 16 daily measurements was collected over a four year period (October 1998-September 2002) from a single air pollution monitoring site, located in a medium-size residential area (see details in [10]) in northern Christchurch City, of air pollution and climate is investigated; PM_{10}, SO_2, CO_2, relative humidity (RH), an indication of the temperature inversion formation (calculated from the difference between the temperatures at 1m and 10m above the ground, with negative values indicating temperature inversion formation, labelled separately as TG, TT and TD), wind speed (WS), wind direction (Wdir), atmospheric pressure (P), radiation hours (Rad), sunshine hours (Sun), rainfall (Rain), maximum and minimum daily temperature and the average of these (TMax, TMin and TAv). All air pollution and climate data were scaled (no specific units). Over the same period, daily counts of acute hospital admissions due to respiratory system problems (ICD-9: 460-519) were obtained for residents domiciled within 2 km (age 0-98 years, n=878 for female, and n=1061 for male) of the air pollution monitoring site. The studied data contained a maximum of about 4% missing values, mainly from SO_2 and temperature inversion data points, but were separately imputed and did not significantly alter results.

3 Results and Discussion

3.1 Key Attributes for the Admission Outbreak

Table 1 shows a summary of numbers and relative reduction (in %, with higher proportions indicating greater reduction) of selected numbers of attributes. TNSU selected only three attributes (minimum subset and 81.3% reduction) whereas CNS selected 11 attributes (maximum subset and 31.3% reduction). TNSU and TNSP selected the smallest numbers of attributes. The rest of the AS methods selected more than 7 out of 16, so about half of the attributes were removed. Table 1 also shows a summary of selected attributes and an assessment of top 3 selected attributes over all AS methods. The TNSU selected TG, CO and RH. The top 3 most commonly

selected attributes over all AS methods are CO and RH, which are ranked highest (4 times) and followed by SO_2 and TG (3 times). It could be said that these four attributes are key factors that can help improving or are underline potential factors of the admission prediction. Additionally, all three attributes selected by TNSU are three of the four most commonly selected attributes by all other AS methods.

3.2 Selection of Attributes for Ant-Miner and C4.5

In Table 2, the means of Ant-Miner and C4.5 from the training sets shows the mean of Ant-Miner (μ=65.8) is not significantly larger, or even equal (p=0.05 for one-tailed using assuming equal variances, as F-test for equal variances shows p=0.13) to C4.5 (μ=65.0). While using all attribute sets recorded similar classification accuracies (Ant-Miner; 64.5% and C4.5; 63.8%), the removal of attribute was not carried out to significantly improve the classification accuracy. However, the proportion of attribute removal was significantly successful (up to 81% for TNSU) and the quality of classification accuracy was maintained, even with much smaller attribute sets.

Although, overall Ant-Miner classification accuracy (training set) recorded slightly higher classification accuracy with the ranking filter approaches; RLF (66.3%), TNSP

Table 1. Numbers of selected attributes, relative attribute reduction (in %) and a summary of selected attributes

Proportion of original class: yes 49% no 51%	TNSU	TNSP	RLF	IG	CFS	CNS	Full	Attribute frequency (Top 3)
# of selected attributes	**3**	5	7	9	8	11	16	
Reduction of attributes (%)	**81.3**	68.8	56.3	43.8	50.0	31.3	0.0	
Ranking of attributes								
1 (most important)	TG	TG	RH	TG	SO_2	SO_2	SO_2	**3**
2	CO	RH	TD	TT	CO	CO	CO	**4**
3	RH	CO	SO_2	Rad	RH	PM_{10}	PM_{10}	1
4		SO_2	TD	Tmin	TG	RH	RH	**4**
5		Tmax	Tav	TD	TD	TG	TG	**3**
6			Rad	Tmax	Rad	TT	TT	1
7			Tmax	Tav	Tmax	TD	TD	1
8				SO_2	Tmin	Rad	WS	
9				RH		Tmax	Wdir	
10						Tmin	P	
11						Tav	Rad	1
12							Sun	
13							Rain	
14							Tmax	
15							Tmin	
16 (the lowest ranking)							Tav	

Table 2. Summary of classification accuracy for training and test sets for Ant-Miner and C4.5 classifiers

Classification	Full (before AS)	TNSU	TNSP	RLF	IG	CFS	CNS	Mean	SD	Two-sample means of Ant-Miner and C4.5
Ant-Miner (training)	64.5	66.0	66.2	**66.3**	66.1	64.8	65.5	65.8	0.6	p=0.05 (one-tail)
C4.5 pruned (training)	63.8	65.8	65.9	65.9	64.0	64.4	63.8	65.0	1.0	
Ant-Miner (test)	54.3	**59.3**	55.7	55.0	47.9	54.3	56.4	54.8	3.8	p=0.37 (one-tail)
C4.5 pruned (test)	54.3	57.9	54.3	55.0	56.4	54.3	54.3	55.4	1.5	
Ant-Miner rules	7	7	7	7	7	8	8			
C4.5 leaves	7	4	7	3	4	8	7			
C4.5 size of tree	13	7	13	5	7	15	13			

(66.2%), IG (66.1%) and TNSU (66.0%), compared with the subset evaluator; CNS (65.5%) and CFS (64.8%). Similarly, overall C4.5 classification accuracy (training set) recorded similar classification accuracies, but also filter AS approaches; RLF and TNSP (65.9%), and TNSU (65.8%) except IG (64.0%) performed slightly better than subset evaluator approaches CFS (64.4%) and CNS (63.8%). Note that Table 2 also shows the information about the constructed rule, e.g., size of tree or number of rules, but are not specifically discussed.

3.3 Testing Prediction Performance with Selected Attributes for Ant-Miner and C4.5

Results of predicted admission accuracy are also summarised in Table 2. The two means of Ant-Miner and C4.5 using the test sets suggests that there is no significant evidence to say that the mean of C4.5 prediction accuracy (μ=55.4%) over all AS methods are higher than or equal to Ant-Miner (μ=54.8%) (p=0.37 for one-tail using assuming unequal variances, as F-test for equal variances shows p=0.03). While the original class proportion of *yes* is 49% and *no* is 51%, it could be said that both classifiers are slightly more effective than just guessing either class, especially TNSU for Ant-Miner, which achieved the highest prediction accuracy around 60%. In fact, TNSU also obtained the highest prediction accuracy for C4.5 (57.9%). On the other hand, IG recorded the lowest prediction accuracy for Ant-Miner (47.9%) and CFS and CNS recorded the lowest for C4.5 (54.3%). As previously mentioned, the AS method may be slightly more effective on Ant-Miner than C4.5.

Even though the classification accuracy was not significantly improved with fewer attributes, a possible reason why TNSU provides the highest prediction accuracy for Ant-Miner over C4.5 can be considered that TNS searches important attributes by assessing the connectivity of adjacent nodes in the decision trees; frequently connected pairs of attributes in the decision tree are more important than ones that are not connected. Ant-Miner rules are constructed by pheromone trails, which produce high solutions based on a high probability pair of attributes during updating pheromone iteratively as ants write, read and estimate the amount of pheromone trail to build a good solution [11]. Hence, the attributes that are selected by TNS may strengthen the search between attributes for Ant-Miner because Ant-Miner similarly searches attributes that have higher probability between nodes (pheromone trial). Providing fewer but specifically selected representative attributes may help increase efficiency in finding a solution in the Ant-Miner in less confusing manners. Whereas IG measures the information gained with respect to class, it may not directly consider the strengths between attribute nodes. Surprisingly, CFS produced the same classification accuracy regardless of removing or full attributes sets for both Ant-Miner (54.3%) and C4.5 (54.5%), even though CFS selects individual features at predicting each class along with the level of inter-correlation [2], which could strength the path that was taken during the Ant-Miner search. The studied data set may not have such good level of inter-correlation.

4 Conclusions

This paper examined six different data mining attribute selection (AS) methods, TNSU, TNSP, RLF, IG, CNS and CFS, to extract key climate and air pollution factors by predicting the acute respiratory admission counts with Ant-Miner and C4.5

algorithms. TNSU preformed best to remove up to 80% of the attributes by selecting only three attributes; temperature at ground level, carbon monoxide and relative humidity, and obtained a classification accuracy improvement (from 2% to 5%) for both Ant-Miner and C4.5. All other AS methods removed approximately half of the attributes, seem to trade off between attribute reduction performance and maintaining prediction accuracy. This is a preliminary experiment using data mining AS methods on environmental and health study with Ant-Miner. It will be expected to keep investigating other newer Ant-Miner algorithms, such as Ant-Miner 2, ACO-Miner, and TACO-Miner, with much larger attribute sets in future.

Acknowledgments. Thanks to Mr. P. Pearson for editorial work and the reviewers for useful comments. Data were provided by ECan, University of Canterbury (Dept. of Physics) and NIWA, and CDHB (URB/06/05/031).

References

1. Fukuda, K., Martin, B.: Decision Trees as Information Source for Attribute Selection. In: SSCI 2009 IEEE CIDM, pp. 101–108 (2009)
2. Hall, M.A., Holmes, G.: Benchmarking Attribute Selection Techniques for Discrete Class Data Mining. IEEE Trans. Knowl. Eng. 15, 1437–1447 (2003)
3. Jensen, R., Shen, Q.: Fuzzy-Rough Sets Assisted Attribute Selection. IEEE Trans. Fuzzy Syst. 15, 73–89 (2007)
4. Witten, I.H., Frank, E.: Data mining: Practical machine learning tools and techniques with Java implementations, 2nd edn. Morgan Kaufmann, San Francisco (2005)
5. Quinlan, J.R.: C4.5: Programs for Machine Learning. Morgan Kaufmann, San Mateo (1993)
6. Kira, K., Rendell, L.: A practical approach to feature selection. In: 9th ICML 1992, pp. 249–256 (1992)
7. Liu, H., Setiono, R.: A probabilistic approach to feature selection: a filter solution. In: Proc. 13th ICML 1996, pp. 319–327 (1996)
8. Parpinelli, R.S., Lopes, H.S., Freitas, A.A.: Data Mining With and Ant Colony Optimization Algorithm. IEEE Transactions on Evolutionary Computing 6, 321–332 (2002)
9. PEHG, Annual summary of outbreaks in New Zealand, Population and Environmental Health Group Institute of Environmental Science and Research Ltd. (2007)
10. Fukuda, K., Takaoka, T.: Analysis of Air Pollution (PM10) and Respiratory Morbidity Rate using K-Maximum Sub-array (2-D) Algorithm. In: SAC 2007, pp.153–157 (2007)
11. Dorigo, M., Stützle, T.: Ant Colony Optimization. MIT Press/Bradford Books, Cambridge (2004)
12. Ke, L., Feng, Z., Ren, Z.: An efficient ant colony optimization approach to attribute reduction in rough set theory. Pat. Rec. Let. 29, 1351–1357 (2008)
13. Pawlak, Z.: Rough sets: theoretical aspects of reasoning about data. Kluwer, Boston (1991)
14. Aghdam, M.H., Ghasem-Aghaee, N., Basiri, M.E.: Test feature selection using ant colony optimization. Expert Sys. Appl. 36, 6843–6853 (2008)
15. Wang, Z., Feng, B.: Classification rule mining with an improved ant colony algorithm. In: Webb, G.I., Yu, X. (eds.) AI 2004. LNCS (LNAI), vol. 3339, pp. 357–367. Springer, Heidelberg (2004)
16. Thangavel, K., Jaganathan, P.: Rule Mining algorithm with a new ant colony optimization algorithm. In: Int'l conf on computational intelligence and multimedia applications, pp. 135–140 (2007)
17. Liu, B., Abbas, H.A., Mckay, B.: Classification rule discovery with ant colony optimization. In: IEEE/WIC Intelligent Agent Technology, pp. 83–88 (2003)