Mining for Spatio-Temporal Distribution Rules of Illegal Dumping from Large Dataset

Bo Fan
Shanghai Jiao Tong University

Long Chen
Shanghai Jiao Tong University

Yih Tng Chong
National University of Singapore

Zhou He
National University of Singapore

Acknowledgement

This research is funded by the Singapore National Research Foundation and the publication is supported under the Campus for Research Excellence and Technological Enterprise (CREATE) programme, the National Science Foundation of China (No. 91024007, 70901052, 71371122), and National Social Science Foundation of China (No.14ZDB152).

Abstract

Illegal dumping has been an issue to be dealt with by the authorities. The incidents distribute across spatial and temporal domains, possibly with recurring patterns.

To assist in addressing the issue, these patterns in the form of classification rules can potentially be mined from large datasets collected by the authorities. This research represents a novel work in discovering rules described by spatio-temporal features of the illegal dumping activities. A feature selection methodology that considers a range of techniques employing differing optimality criteria is proposed. A hybrid algorithm is developed by combining the proposed method to the C4.5 algorithm. A series of experiments demonstrated the advantages of the proposed algorithm. The feature selection approach is shown to balance the different optimality criteria, overcoming the dominance of any individual criteria. This work further shows that the generated spatio-temporal rules, when generated and implemented in information systems, are potentially applicable in preventive and enforcement work by the authorities.

K**eywords:** Illegal dumping, rule mining, feature selection, large data, C4.5

ACM Categories: H.4.2

Introduction

Illegal dumping in cities has been a major issue for municipal offices. It affects the image, environment and quality of life of residents in cities. Data provided by the Shanghai environmental office shows that there were more than 5800 cases within 115 km2 of land area in the city. Every year in Shanghai, millions of dollars have been spent on cleaning up illegally discarded waste. To address the issue, characteristics of the cases in terms of location and timing should be understood. Such information will facilitate the monitoring, enforcement and clearing of the waste. According to recent research, illegal dumpling incident is determined not only by human behavior, but also by the surrounding demographic factors and rubbish recycling facilities (Barr et al. 2007; De Young 1988). It has been known that spatial factors can be associated to the occurrences of events (Fan and Luo. 2013).

This research is based on a large dataset that recorded the occurrences of illegal dumping in the time and space domains, and for each case, the surrounding geographical features. Based on the dataset, this research aims to extract classification rules that describe the spatio-temporal distribution patterns of the cases. With the knowledge of the patterns, planning of waste management can be better performed by the municipal offices. In this work, to solely apply classification techniques will result in rules with redundant features (Quinlan, 1993); a feature selection process is required to derive an optimal subset of features prior to the application of a classification technique. This approach will increase the computational efficiency and prevent the derivation of sub-optimal rule set due to the presence of redundant features. This work proposed a novel methodology to address the problem of mining the large data described above, based on the combination of a proposed feature selection approach to the C4.5 algorithm. A series of experiments demonstrated the advantages of our proposed algorithm.

In this paper, related research on waste generation and rule mining methods is summarized in Section 2. In Section 3, the data treatment process is illustrated, which includes data description and discretization. A novel spatio-temporal distribution rule mining method that incorporates feather selection mechanism into C4.5 is proposed. Several experiments are conducted to test the efficiency of the algorithm. Finally, rules with practical significances are identified, and their managerial implications with relevance to information systems are discussed in Section 4. A summary, limitation of this work and further research direction are presented in Section 5.

Literature Overview

Factors affecting waste generation

Research in waste management systems not only involves waste generation, transportation, processing and the corresponding social and environmental issues, it also requires the studies of the technologies, economics, sociology and political sciences (Weberman, 1980). McDougall et al (2008) applied a lifecycle approach in waste management for the utilization, neutralization and reduction of waste. Traditional studies in waste generation factors and rate prediction were mainly based on population statistics and social factors, utilizing such techniques as static modeling, geometric mean, exponential smoothing, and single or multivariate linear regression methods. Based on static factors such as population size, income level and area of housing units, Grossman et al. (1974) and Niessen and Alsobrook (1972) applied linear regression model to predict the average rate and patterns of waste generation. Chang and Lin (1997) applied econometric forecasting models with considerations various dynamic factors in waste recycling to address the uncertainties in the volume of waste collected. The above-mentioned methods utilized static predictive models, requiring complete social and environmental data to perform predictions. Further, the results do not reflect the dynamic nature of the waste generation. As studies based on incomplete data is common in practice, Deng et al (1982) proposed the application of grey dynamic model to address the issue of missing data. Grey models generate imprecise prediction based on differential predictive models, when only small amount and missing data are presented. Chen and Chang (2000) applied grey models to forecast waste generation with reasonable bounds of accuracy. Subsequently, a research applied system dynamics modeling to predict waste volume in rapidly developing cities, based on limited amount of data; a case of Texas in the US was used to validate the performance of the approach (Dyson and Chang 2005). Mashayekhi (1993) likewise employed system dynamics modeling in a study of waste transfer system in New York. Sudhir et al. (1997) advanced the studies using system dynamics to investigate the characteristics of interactions between system components within a municipal solid waste management system. Hornik et al.(1995) surveyed sixty-seven studies to extract factors affecting waste generation, and classified them into four types – internal motivation, external motivation, internal condition, and external condition.

Based on independent variables related to the environment, Taylor and Todd (1995) employed behavioral planning management to study the effects of the variables on recycling behaviors and against illegal dumping. Barr et al. (2003) studied the attitudes of a group of residents towards recycling household waste. De Young (1988) explored differences between recyclers and non-recyclers and found that information levels received by the two groups had vast disparity. Based on the case of a Spanish city, Grace et al. (2002), analyzed the factors affecting city-dwellers' recycling behavior and resisting illegal dumping; it was revealed that household income was inversely related to recycling behavior, and that age had a positive relationship to recycling behavior.

Data classification and feature selection

Based on a review of literature in data mining, studies in illegal dumping have not been extensive. Spatial studies mainly addressed the problem of waste collection and transportation Tasaki et.al (2007). This study obtained a dataset that describes the cases of illegal dumping incidents and the related GIS information. To exploit the available data, this research proposed a novel method that is based primarily on the C4.5 algorithm with a combination of multiple feature selection techniques, in order to increase its efficiency and interpretability of the generated rules.

There is a range of data mining methods based on classification modeling, including artificial neural network, K-nearest neighbor algorithm, SVM (Support Vector Machine), and decision tree methodologies. Based on the study of Kotsiantis et al.(2006) on the various methods' applicability and performances, the decision tree approach was found to be better overall. Firstly, the structure of a decision tree is relatively simple and can be directly interpreted by human. Secondly, decision tree models tend to have higher efficiency, which makes them suitable for machine learning from large dataset. Next, no knowledge is required beyond the assigned training dataset, which is suitable for addressing non-numerical data types. Further, decision trees usually have higher data analytics ability. The data obtained for this research is relatively large, which involves various qualitative features, and the relationships between features are not explicit. Due to that, this work applied the decision tree approach. C4.5 is an embedded method that inherently incorporated feature selection function based on the entropy principle (Dai and Xu 2013), which is however not strong in removing redundant features. Furthermore, the feature selection function of C4.5 involves repeated scanning and reordering, thus the efficiency will be reduced when the training dataset and feature space is large.

Feature selection is a necessary step in data mining. Raw data obtained for study may contain an extensive number of features of which many may be redundant and irrelevant. These redundant features are likely to negatively affect the efficiency and results of the data mining process. In the fields of machine learning and pattern recognition, researchers have differing opinions toward the definition of feature selection. Dash et al.(2000) and Fan and Zhang (2009) opined that a small size of attribute set should be chosen if the distribution and accuracy of the data classification algorithm cannot be improved. John et al.(1994) and Liu and Yu (2005)highlighted that feature selection is to increase classification accuracy on the premise that the data classification accuracy being adequate. We can therefore see that different approaches of feature selection will produce different outcomes. As there is no single method that is optimal for all types of data, the choice of specific or combinatorial feature selection methods should be context-driven based on the characteristics of the subjected data.

Research gaps

The research gaps identified in this work are on two levels – domain problem and methodological – as described in this section.

Existing research in illegal dumping focuses on the management and strategies of surveillance and control. For example, Tasaki et.al (2007) examined and evaluated occurrences and sizes of illegal dumping to discover potential dumping sites using GIS (Geographic Information System) technology. Ichinose et.al （2011） examined the relationships between the provision of waste treatment facilities and the frequency of illegal dumping. Further, Tam et.al (2014) simulated the effects of landfill charges and penalties from illegal dumping in Shenzhen, China by employing system dynamics method. Behavioral studies in waste management have emphasized on waste sorting at source for recycling. More behavioral studies on the disposal of typical domestic waste have been done, compared to studies on illegal dumping activities of larger scales. The related studies reviewed above cover a range of approaches. The majority of them applied economic analysis, statistics, regression analysis, system dynamics and grey models to analyze the factors of waste generation. Based on our research, the problem of illegal waste dumping has not been investigated using the data mining approach in the past.

Noting that the behavior of waste dumping is affected by both timing and geographical factors, this research implemented a data mining approach to discover spatio-temporal patterns of illegal waste dumping. Through extracting knowledge from large data, this research aims to contribute with an approach that provides new insights to problems in the domain of illegal waste dumping.

On the methodological level, current research in feature selection has emphasized on the enhancement and optimization of the methods' efficiencies and performances. Optimality criteria of the available methods vary and their applications on datasets of differing characteristics will produce different results. In practice, an applicant would not know on the outset the specific feature selection techniques that would be more appropriate with respect to a particular dataset obtained for analysis.

To this end, this research proposed a novel methodology that assesses the suitability of a set of feature selection techniques that may apply differing optimality criteria, with reference to a specific dataset acquired for analysis. The resulted optimal feature subset consists of features that performed well across the set of features selection techniques. This approach of feature selection is able to balance the different optimality criteria adopted by the various techniques, overcoming the dominance of any individual criteria in the process. This work is based on the well-known C4.5 classification rules mining technique. C4.5 is an embedded method incorporating feature selection functions, which is solely based on the ratio of normalized information gain as its effective feature selection criterion. This research applied a novel step that integrally considers a set of optimality criteria, which is combined with C4.5 method as mentioned above. This approach is able to increase the feature selection capability of the C4.5 technique in remove redundant features, and with a reduction in features, the computation efficiency is expected to increase. The evaluation of the proposed methodology will compare the capability of filtering redundant features, computation speed, computation complexity and the structure of decision trees.

In addressing the two research gaps discussed above, this work contributes in terms of (1) pioneering the application of the data mining approach on real world datasets and spatial GIS information, in providing spatio-temporal insights for illegal waste dumping management, and (2) a proposed methodology that improves the feature selection capability of C4.5, which is critical in real world problems where features are often high in number, for example, in the GIS information employed in this study.

Spatio-temporal rule mining method

Data treatment

A dataset acquired for this research was sourced from the backend data center of a Shanghai Digital City Management Office. The data was captured from 10 October 2008 to 31 March 2013, recording a total of 26450 cases. In the raw data, there are 9 features, namely time, task ID， street， address, problem description, X-coordinate, Y-coordinate, department responsible, and time of case closure. The second dataset was sourced from a geographical database department of the Shanghai Digital City Management Office. Based on the relevant non-confidential layers of GIS-based map acquired, a spatial search was performed around every geographical point of the recorded illegal dumping, within the radius 1000 metres. The spatial features of every layer were tabled, which later served as the features hypothesized to be associated to the recorded illegal dumping activities. Upon analysis, 18 types of spatial features were resulted.

Further, data cleansing, transformation, assembling and discretization were performed. Based on the X and Y coordinates data of the illegal dumping cases, this study found 5 data points that fit outlier characteristics. They were therefore removed from the dataset to ensure proper data discretization later in the process. Assembling of datasets is based on the common ‘task number’ from the two datasets. The field recording textual descriptions of the illegal dumping cases were transformed to a field representing the different types of waste involved, namely construction waste, domestic waste, or other waste type. In this work, the generation of decision trees using the C4.5 method uses information entropy measure. In this case, if the data discretization adopted here is frequency-based, the use of information entropy would not be applicable. As such, distance-based data discretization is adopted in this work.

Feature selection

This research applied 5 different feature optimality measurements on 18 features, to more robustly assess the importance of each spatial feature. Ranking search methods and the FCBF methods were used to rank the features(Liu and Yu 2005). Features that commonly appear in the resulted top ten ranking under all 5 types of measurement were selected as the optimal feature subset. The results of the ranking are as shown in Table 1.

The five different types of measurements are as described below.

Pearson’s Correlation has been widely applied to measure the strength of linear correlation between two variables. It was developed based on the idea of Francis Galton in the 1980s. Equation 1 shows the correlation between two random variables X and Y.

$$\rho\text{=}C\mathrm{or}(X,Y)=\frac{Cov(\mathrm{X},\mathrm{Y})}{\sqrt{Var(X)Var(Y)}} \quad （1）$$

Cov(X,Y) represents the covariance between X and Y, while Var(X) and Var(Y) represent the variance of X and Y respectively. When the correlation is at the value of 1, the relationship between X and Y can be expressed as $Y = aX + b$, where a>0; when the correlation is at the value of -1, the relationship between X and Y can be expressed as $Y = aX + b$, where a<0. If X and Y are independent to each other, the value of correlation will be 0.

Information gain is based on information entropy theory. The entropy of a random variable is a measure of the uncertainty of the random variable. The entropy of random variable X is as shown in Equation 2 (Quinlan 1993).

$$H(X) = -\sum_i P(\mathrm{x}_i)\log_2(P(\mathrm{x}_i)) \quad （2）$$

The entropy of X with the knowledge of Y is as shown in Equation 3.

$$H(X|Y) = -\sum_j P(y_j)\sum_i P(x_i|y_i)\log_2(P(x_i|y_i)) \quad （3）$$

Table 1. Results of Ranking by Feature Selection Methods

	Symmetrical Uncertainty	Pearson's Correlation	Gain Ratio	Information Gain	OneR
1	bus station	bus station	bus station	bus station	cross junction
2	community	community	community	community	community
3	cross junction	cross junction	cross junction	cross junction	bus station
4	road	long distance vehicle stop	long distance vehicle stop	road	road
5	car park	park	road	car park	car park
6	building	government building	building	building	long distance vehicle stop
7	long distance vehicle stop	building	hotel	Hotel	residential area
8	hotel	car park	Government building	long distance vehicle stop	park
9	government building	road	park	government building	government building
10	park	hotel	Car park	Park	school
11	school	residential area	taxi office	residential area	building
12	residential area	taxi office	school	school	hotel
13	taxi office	customer address	residential area	taxi office	mile stone of high bridge
14	mile stone of high bridge	school	mile stone of high bridge	mile stone of high bridge	Customer address
15	customer address	hospital	Customer address	Customer address	taxi office
16	hospital	Tourist attraction	hospital	hospital	hospital
17	tourist attraction	mile stone of high bridge	tourist attraction	tourist attraction	river
18	river	river	river	river	tourist attraction

The above P(Xi) as the priori probability of the random variable X, while P(Xi|Yi) is the conditional probability of X under the observation of random variable Y (Mukras et al. 2007). Information gain (IG) is defined by Equation 4.

$$IG(X|Y) = H(X) - H(X|Y) \quad (4)$$

The importance of a feature is based on the information gain or entropy increase in the classification system brought about by the inclusion of the feature; the higher the information gain, the more important is the feature.

Gain Ratio leads to the concept of intrinsic information. The importance of a feature will be reduced with the increase in its intrinsic information. Gain ratio is an approach of compensation to address the problem of information gain. However it could also lead to over-compensation, leading to the selection of features with low intrinsic information. Gain ratio is the ratio of information gain and intrinsic information, see equation 5 (Dai and Xu 2013).

$$G\mathrm{ainRatio}(\mathrm{X,Y}) = \frac{\mathrm{IG}(X|\mathrm{Y})}{IntrinsicInfo(X,\mathrm{Y})} = -\frac{\mathrm{IG}(X|\mathrm{Y})}{\sum \frac{|X_i|}{X}\log_2\frac{|X_i|}{X}}$$

（5）

Symmetrical uncertainty is the measurement of correlation between features, first developed by Press et al. (1992).Based on the concept of information gain, symmetrical uncertainty avoids the bias of information gain, which leads to an inclination to select the attribute with more values. It can overcome the shortcoming of linear correlation measurement, where problems are non-linear. The formula is as shown in Equation 6.

$$SU(X,Y) = 2[\frac{IG(\mathrm{X}|\mathrm{Y})}{H(X)+H(\mathrm{Y})}] \quad （6）$$

OneR classifier produces simple rules based on one feature only. For each feature in the training dataset, OneR builds a rule, and selects the rule with the smallest error (Witten and Frank 2005).

Based top the 10 spatial features selected by each of the five feature selection methods (as shown in Table1), 8 spatial features commonly selected across the methods were identified from the total of 18 features. They are bus station, community centre, cross junction, road, car park, long distance vehicle stop, park and government building.

A comparison between C4.5 decision tree and the proposed method

C4.5 decision tree algorithm is based on ID3 (Quinlan 1993). In WEKA, a data mining application tool, the J48 method is an implementation of C4.5. This work applied J48 for the mining of classification rules.

Especially, the parameter 'minNumObj' within the J48 algorithm is the minimal incidents, which is the threshold of incident number each node holds. This parameter is adjusted in this study to vary the granularity of pruning the decision tree. It functions to control the number of nodes and complexity of the tree, which has effect on the interpretability of the resulted decision tree. If there are too many nodes, the rules will be longer and will have lower interpretability; beside, over-fitting will be resulted. For these reasons, a limit to the number of nodes has to be set to ensure the readability and applicability of the rules. Upon feature selection, the dataset has 21 features and 26444 instances, which was subjected to the mining of classification rules by the J48 method. Based on repeated experimentation and analysis, with the parameter 'minNumObj' set at 500, the results of tree pruning led to stable classification accuracy. Further, the resulted decision tree has good interpretability, providing realistic value and actionable information for applications in the management of illegal dumping.

A comparison of performances between the original C4.5 and C4.5 combined with the proposed feature selection method was done in this study. This experiment employed a number of instances(NI), the threshold instances of the leave node(minNumObj) and control variables. The performances of the two approaches were observed and analyzed from numerous perspectives. This work applied a random sampling function in the data pretreatment module of the WEKA platform. Based on the dataset both before and after feature selection, with 26444 instances, starting from 1000 and ending at 26000 with an increment of 1000 instance for every step, 26 cases were sampled as training data. The minimum number of nodes is controlled via the 'minNumObj' parameter, at 400, 500 and 600. The 26 training instances were applied to the two competing methods, and the result of the experiment with minNumObj=500 is shown as an example in Table 2.

Computation time analysis

In applying the original C4.5 algorithm, the computation complexity is O(N*Mlog2M), where N is the number of feature and M is the number of instances. Upon combining the proposed feature selection technique, where n is the number of feature after the selection process (n<N), the computation complexity of the feature selection process has to be considered, other than that of the C4.5. Within the feature selection process, two processes have to be considered. Firstly, the time complexity of feature assessment, related to the M number of instances, is O(M). Secondly, the process of features generation has the time complexity ranging from the best case O(n) to the worst case O(n2). Considering the worst case of the process, the proposed approach has a complexity of O(n2)+ O(M)+ O(n*Mlog2M). In the studies of data mining, the feasibility is high when the number of instances is far higher than the number of features, as in the case of this research. In this work, M is higher than n by 2 to 3 orders (M>>n), as such, O(n2)+ O(M) in relation to O(n*Mlog2M) is negligible. As a result, the gap of computation complexity between the two approaches is the difference between O(N*Mlog2M) and O(n*Mlog2M). As n<N, the proposed approach will be relatively less complex in theory. Through experimentations, regardless of the three values of minNumObj, the proposed approach will require less computation time. With the increase in the number of instances, the computation time gap between the two approaches will increase. Figures 1a, 1b, 1c and 1d show the computation time with minNumObj=300, 400, 500 and 600 respectively.

Table 2. Result of experiment with’minNumObj=500’

	Original C4.5					C4.5 with the combination of the proposed feature selection method			
NI	T	A	LN	TN	RN	T	A	LN	TN
1000	0.002	40.94	1	1	0	0.002	40.94	1	1
2000	0.01	53.39	5	7	1	0.003	53.68	4	6
3000	0.02	56.4	5	8	1	0.01	56.98	4	6
4000	0.02	56.82	5	7	0	0.01	56.76	5	7
5000	0.02	58.18	4	6	0	0.02	58.21	4	6
6000	0.04	57.85	4	6	0	0.02	58.63	4	6
7000	0.03	58.54	9	14	1	0.01	58.58	7	10
8000	0.03	58.36	9	14	1	0.02	58.26	9	13
9000	0.03	58.46	7	10	0	0.03	58.4	7	10
10000	0.05	58.59	7	10	0	0.04	58.61	7	10
11000	0.06	58.61	7	10	0	0.04	58.71	7	10
12000	0.07	58.64	9	14	1	0.07	58.93	10	15
13000	0.08	59.47	11	18	2	0.06	59.6	8	12
14000	0.08	59.65	8	12	0	0.08	59.71	8	12
15000	0.09	59.63	11	18	2	0.05	59.44	8	12
16000	0.09	59.17	11	18	2	0.08	59.19	13	21
17000	0.11	59.8	11	18	2	0.08	59.58	8	12
18000	0.18	59.72	13	22	3	0.07	59.48	10	16
19000	0.19	59.76	13	22	2	0.08	59.31	14	23
20000	0.19	59.22	10	16	1	0.11	59.18	10	16
21000	0.15	59.81	13	22	3	0.11	59.44	17	28
22000	0.19	59.54	12	20	2	0.1	59.16	16	26
23000	0.22	59.75	14	22	2	0.1	59.42	14	22
24000	0.2	59.47	20	33	4	0.11	59.43	11	17
25000	0.34	60.07	15	25	3	0.12	59.97	15	24
26000	0.35	60.12	15	25	3	0.14	59.95	15	24
26445	0.34	59.72	13	22	4	0.19	59.96	10	16

NI: number of instances; T: computation time; A: accuracy; LN: leave nodes; TN: tree nodes; RN: redundant attribute number

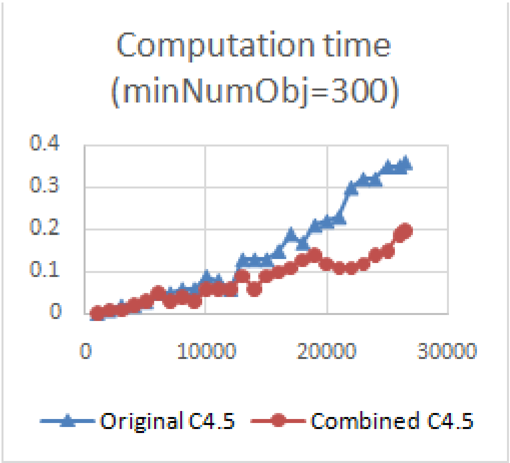

Figure 1a. Computation time (MinNumObj=300)

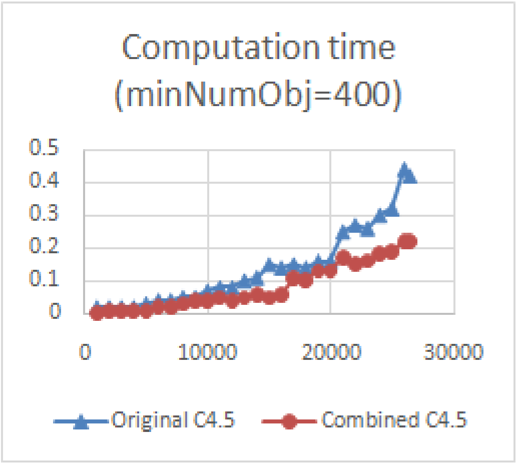

Figure 1b. Computation time (MinNumObj=400)

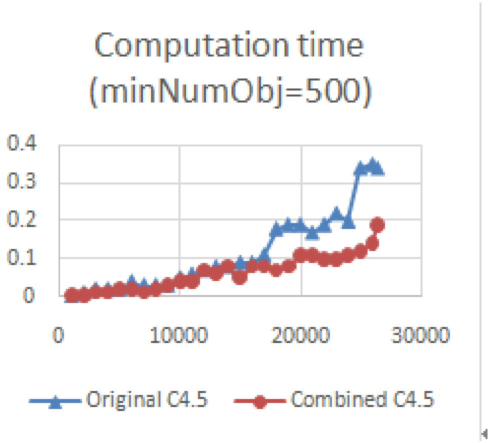

Figure 1c. Computation time (MinNumObj=500)

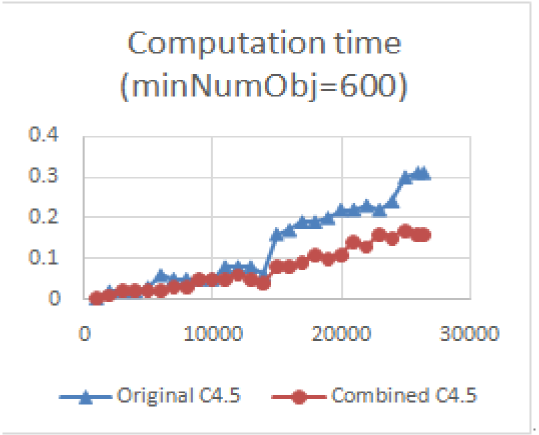

Figure 1d. Computation time (MinNumObj=600)

Accuracy analysis

Based on the result of the experiments, it can be seen that the accuracy of the classification rules will increase with the number of instances. It shows that C4.5 requires a certain number of training instances to effectively generate classification rules. However, it has been also noted that there is a leveling of accuracy when the number of training instance is increased. There was no marked increase of accuracy when the proposed method is compared to the original C4.5.

Tree nodes and leaf nodes

The number of tree nodes and leave nodes are a measurement of the complexity of the resulted decision tree. The experiments show that the number of nodes as well as the accuracy of the classifiers will increase with the increment of training instances. However, with a large number of nodes, the decision tree will appear complicated, leading to low readability, interpretability and applicability in decision-makings. In this case, the decision tree is likely to have over-fitted the problem. Based on experimentation and experts' opinion on the results, it was found that a balance between accuracy and interpretability could be reached with the parameter 'minNumObj' set at the value of 500.

Number of redundant features

Redundant features refer to the 10 features that were removed based on the 5 feature selection techniques adopted in the proposed methods. In the case of using the original C4.5 approach, some of the redundant features formed part of the decision variables in the resulted decision tree. The application of only one measurement of feature optimality in C4.5, i.e. the information entropy, is a weakness of the approach in terms of feature selection. With the use of the proposed approach in feature selection, the chances of redundant features appearing in the final decision tree will be greatly reduced.

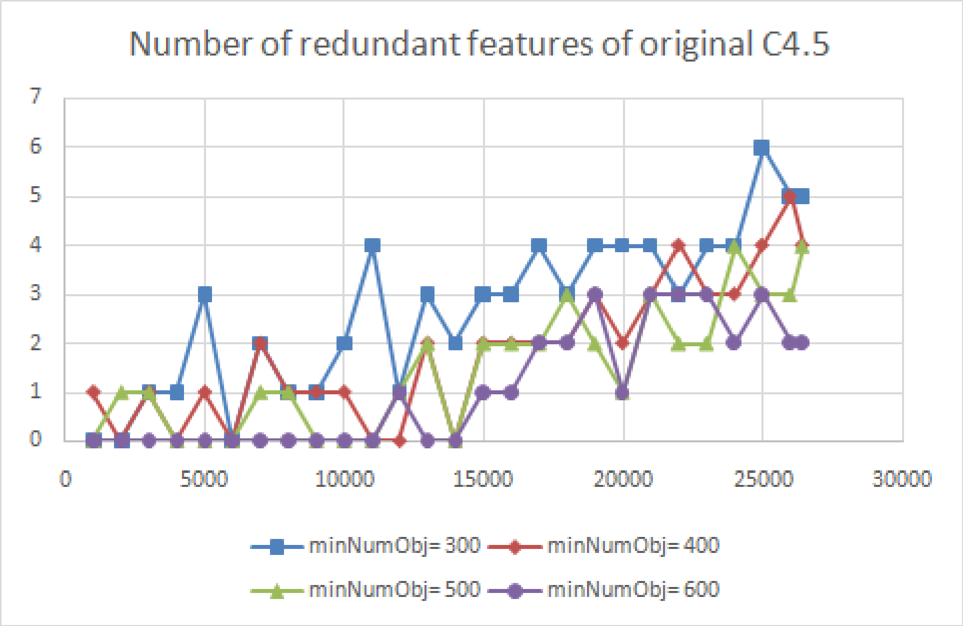

Figure 2. Number of redundant features of original C4.5

Fig.2 shows the appearance of redundant features under the different setting of parameter 'minNumObj' when the original C4.5 method was adopted. Based on analysis in this study, lower number of redundant features appeared when the number of training instances is low. This is due to that the features that led to higher information gain also performed well under other measurement criteria. However, with the increase in the number of training instances and therefore the accuracy of the classifiers, more redundant features were included into the classification system leading to an impact on the quality of the outcome.

Generation and interpretation of the rules

Decision tree and rules evaluation

The decision tree resulted from the proposed methodology has 6 decision attributes, 5 spatial features (bus station, cross-junction, car park, long distance vehicle stop, and government building), and a temporal feature (8:00 – 10:00).There are a total of 10 leave nodes, indicating the generation of 10 classification rules. The rules represent the associations between different types of illegal dumping and spatio-temporal features. The rules can be presented in terms of IF-THEN statements as shown below.

There are 7 rules for domestic waste:

（1）IF ***Bus-station Number =High***，THEN ***Waste type =Domestic waste***;

（2）IF ***Bus-station number =Low*** AND ***Car-park number =High***，THEN ***Waste type = Domestic waste***;

（3）IF ***Bus-station number =Low*** AND ***Car-park number = Medium***，THEN ***Waste type = Domestic waste***;

（4）IF ***Bus-station number = Medium*** AND ***Cross-junction =Low***，THEN ***Waste type = Domestic waste***;

（5）IF ***Bus-station number = Medium*** AND ***Cross-junction = Medium***，THEN ***Waste type = Domestic waste***;

（6）IF ***Bus-station number = Medium*** AND ***Cross-junction = high*** AND ***Long distance vehicle stop = YES***，THEN ***Waste type = Domestic waste***;

（7）IF ***Bus-station number = Medium*** AND ***Cross-junction = High*** AND ***Long distance vehicle stop = NO***AND ***Government building = Yes*** AND ***8:00-10:00= YES***，THEN ***Waste type = Domestic waste***;

There are 2 rules for construction waste:

（8）IF ***Bus-station number = Medium*** AND ***Cross-junction = High*** AND ***Long distance vehicle stop = NO*** AND ***Government building = NO*** AND ***8:00-10:00 = YES***，THEN ***Waste type = Construction waste***;

（9）IF ***Bus-station number = Medium*** AND ***Cross-junction = High*** AND ***Long distance vehicle stop = NO*** AND ***8:00-10:00 = NO***，THEN ***Waste type = Construction waste***;

There is 1 rule for other waste:

（10）IF ***Bus-station number = Low*** AND ***Car-park number = Low***，THEN ***Waste type = Other waste***;

Delphi method is employed to conduct the analysis process. It is a structured communication technique, originally developed as a systematic, interactive forecasting method which relies on a panel of experts (Dalkey and Helmer 1963). Three experts in the field of waste management were invited in this evaluation. The experts were requested to answer questionnaires in two or more rounds. After each round, a facilitator provides an anonymous summary of the experts' forecasts as well as the reasons they provided for their judgments. The experts were encouraged to revise their earlier answers in the light of the replies of other experts. It is believed that during this process, the range of the answers will decrease and converge towards the "correct" answer. After three rounds of evaluation, five rules (i.e. (2), (3), (6), (8) and (9)) are firmly opined as meaningful.

Rules interpretation

Based on the classification rules generated above, it can be seen that spatio-temporal features that are associated with illegal dumping activities are mainly various types of transportation infrastructures, e.g. bus stations, cross-junction and long distance vehicle stops. With an aggregation of these features, there was a higher chance of illegal dumping occurring. This could be due to the lack of or the placement of dumping facilities in these areas coupled with the high vehicular and human traffic.

As far as the sixth rule (6) is concerned, it indicates that illegal dumping had more often occurred at long distance vehicle stops. These locations usually have high human traffic from different parts of the country, and in general, the travelers are not environmentally conscious. Even if there were allocated dumping facilities, the travelers might not use them, leading to the accumulation of waste in these areas. The second (2) and third rule (3) indicated that domestic waste of illegal dumping was often found at places where there are medium to large numbers of car parks. Other than the reason of higher human traffic, it could be observed that there were few waste collection facilities around car parks leading to the illegal disposal of waste. The eighth (8) and ninth rule (9) reflect the patterns of spatio-temporal features in cases of construction waste dumping. Other than features related to transportation infrastructures as discussed above, two other features are government buildings and the '8:00hrs to 10:00hrs' timing. The ninth rule indicates that illegal dumping of construction waste were usually done beyond the mentioned time slot, which is the morning rush hours when the illegal dumping could be spotted by the public. For example, the offenders could choose to commit the act during late hours when human traffic is low. The eighth rule suggests that even if illegal dumping activities were reported between the morning rush hours, it is usually not in an area near to a government building. Government-related premises in proximity could therefore be a deterrent to the act. The illegal dumping of domestic waste is usually of small scale and not obvious in comparison to the illegal dumping of construction waste, which might explain why the illegal dumping of domestic waste having no apparent association to government buildings in proximity.

Managerial implications

Through this research, 'real world' rules that describe the spatio-temporal patterns of illegal waste dumping were discovered. Based on the discovered rules or patterns, four implications are discussed in this section. Further and deeper implication studies on various types of information systems can be performed beyond the scope of this work. For example, a further study on optimizing the planning of waste collection can be done, based on GIS information and the discovered rules.

Optimizing waste collection facilities

The analysis of rules in the preceding part shows that waste disposal cases that are common at transportation infrastructure such as cross-junctions and car parks were likely due to the high density of human traffic coupled with sub-optimal allocation and placement of waste collection facilities. In view of this, it is recommended to increase the number of waste collection facilities and optimize their geographical distribution. Further, in peak periods, collection points and their clearing frequency should be increased. The design of waste collection bins should also be considered. It has been observed that bins that have lids operated by a foot pedal are often not well used, and that metallic bins were often stolen. From these perspectives, design of the bins can be improved to help address the issues.

Improving waste collection planning – from reactive to proactive

In areas with higher rate of waste dumping activities, the waste collection approach commonly involves collection from trash bins. In the current situation, bulks of waste are often seen outside business establishments and around the allocated trash bins. It is recommended that the current reactive-based approach be improved to a proactive one. In other words, waste collection agency may assign appropriately sized waste bins for businesses and residential units and actively clear them via waste collection trucks based on an appropriate schedule. It is further suggested that the use of trash bags be implemented, which can reduce the leakage of waste water and foul odor into the environment.

Implementing regulation at affected location and enforcement during peak periods

The generated classification rules indicated that cases of illegal waste disposal were associated to certain time spans and locations. As such, enforcement officers can be strategically deployed based on the timings and locations indicated by the results of this study. Surveillance cameras can also be accordingly installed. An increase of the penalty for infringement of the related regulation, as a deterrent, is likely to help address the issue.

Refining the quality of the data collected

This study found that the data currently collected by the authority through form-filling processes consists of missing and imprecise information. In the effort of digitizing the city's operation and management, the quality of the data collected is a critical factor. It is recommended that the entire process of data collection be governed by clear and strict protocols, to improve the accuracy and completeness of the collected data. Consistent process audits and reviews will also be beneficial.

Conclusion and future research

This research is driven by the purpose of analyzing the large dataset acquired from the Shanghai government agency, which was expected to help in addressing the problem of illegal dumping. It applied a data mining methodology to uncover classification rules that describe the association of spatio-temporal features to illegal dumping activities. In practice, this study supported the authority in the decision-making processes of illegal dumping management. In research terms, this work contributes through proposing a feature selection methodology that accounts for a range of optimality criteria, leading to a feature subset that combines the results of a set of feature selection techniques. When the proposed methodology is combined to the original C4.5 technique, it was shown to overcome the weakness of C4.5 in feature selection, resulting in higher computation efficiency and a better resultant decision tree through the omission of redundant features.

The research has faced various limitations, which may be addressed in future studies. Due to various administrative and practical reasons, the data received for this study was limited and could be of a wider scope to provide further instructive guidelines for the authority. Firstly, there was some loss of details in the geographical data recording spatial features. Secondly, further types of data such as geographical-based population density, income levels, education levels and housing types will potentially add value to the results of this study. The C4.5 technique applied in this research led to association rules that do not explicate the degree of association between spatio-temporal features and the illegal dumping activities. Further research will address the above limitations and other advanced analysis.

References

Barr S, Ford N.J, and Gilg A.W.B. (2003). Attitudes towards recycling household waste in Exeter, Devon: quantitative and qualitative approaches[J]. Local Environment, Vo. 8, No. 4: pp. 407-421.

Chang N.B, Pan Y.C, Huang S.D. (1993). Time series forecasting of solid waste generation[J]. Journal of Resource Management Technology, Vo. 21, No. 1: pp. 1-10.

Chang N.B, Lin Y.T. (1997). An analysis of recycling impacts on solid waste generation by time series intervention modeling[J]. Resources, Conservation and Recycling, Vo. 19, No. 3: pp. 165-186.

Chen H.W, Chang N.B.(2000). Prediction analysis of solid waste generation based on grey fuzzy dynamic modeling[J]. Resources, Conservation and Recycling, Vo. 29, No. 1: pp. 1-18.

Dai J, Xu Q. (2013). Attribute selection based on information gain ratio in fuzzy rough set theory with application to tumor classification[J]. Applied Soft Computing, Vo. 13, No. 1: pp. 211-221.

Dash M, Liu H.(1997). Feature selection for classification[J]. Intelligent data analysis, Vo. 1, No. 3: pp. 131-156.

Dalkey N, Helmer O. (1963). An Experimental Application of the Delphi Method to the use of experts[J]. Management Science, Vo. 9, No. 3: pp. 458-467.

Deng J.L.(1982). Control problems of grey systems[J]. Systems & Control Letters, Vo. 1, No. 5: pp. 288-294.

Dyson B, Chang N.B.(2005). Forecasting municipal solid waste generation in a fast-growing urban region with system dynamics modeling[J]. Waste Management, Vo. 25, No.7: pp. 669-679.

De Young R.(1988). Exploring the difference between recyclers and non-recyclers: The role of information[J]. Journal of Environmental Systems, Vo. 18, No. 4: pp. 341-351.

Fan B, Zhang P.Z. (2009). Spatially enabled customer segmentation using a data classification method with uncertain predicates[J]. Decision Support Systems. Vo. 41, No. 3: pp. 176–193.

Garces C, Lafuente A, Pedraja M, et al.(2002). Urban waste recycling behavior: antecedents of participation in a selective collection program[J]. Environmental Management, Vo. 30, No. 3: pp. 378-390.

Grossman D, Hudson J.F, Marks D.H.(1974). Waste generation models for solid waste collection[J]. Journal of the Environmental Engineering Division, Vo. 100, No. 6: pp. 1219-1230.

Hornik J, Cherian J, Madansky M, et al. (1995) Determinants of recycling behavior: A synthesis of research results[J]. The Journal of Socio-Economics, Vo. 24, No. 1: pp. 105-127.

Ishii K, Furuichi T, and Nagao Y. (2013). A needs analysis method for land-use planning of illegal dumping sites: A case study in Aomori–Iwate, Japan[J]. Waste Management, Vo. 33, No. 2: pp. 445-455.

Ichinose D, Yamamoto M. (2011). On the relationship between the provision of waste management service and illegal dumping[J]. Resource and Energy Economics, Vo. 33, No. 1: pp. 79-93.

Kotsiantis S.B, Zaharakis I.D, Pintelas P.E. (2006). Supervised machine learning: A review of classification techniques[J]. Artificial Intelligence Review, Vo. 26, No. 1: pp. 159–190.

John G.H, Kohavi R, Pfleger K.(1994). Irrelevant Features and the Subset Selection Problem. Machine Learning: Proceedings of The Eleventh International.

Liu H, Yu L. (2005). Toward integrating feature selection algorithms for classification and clustering[J] .IEEE Transactions on Knowledge and Data Engineering, Vo. 17, No. 4: pp. 491-502.

Mashayekhi A.N. (1993). Transition in the New York State solid waste system: a dynamic analysis[J]. System Dynamics Review, Vo. 9, No. 1: pp. 23-47.

McDougall F.R, White P.R, Franke M, et al. (2008). Integrated solid waste management: a life cycle inventory[M]. John Wiley & Sons.

Melosi M.V. (1981). Garbage in the cities: refuse, reform, and the environment[M]. University of Pittsburgh Press.

Mukras R, Wiratunga N, Lothian R, et al. (2007) Information gain feature selection for ordinal text classification using probability re-distribution[C]//Proceedings of the Text link workshop at IJCAI.

Niessen W.R, Alsobrook A.F. (1972). Municipal and industrial refuse: composition and rates[C].Proceedings of national waste processing conference.

Press W.H., Teukolsky S.A, Vetterling W.T, and Flannery B.P. (1992). Numerical Recipes in C[M]. Cambridge University Press, Cambridge.

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

Sudhir V, Srinivasan G, Muraleedharan V.R.(1997). Planning for sustainable solid waste management in urban India[J]. System Dynamics Review, Vo. 13, No. 3: pp. 223-246.

Seror N, Hareli S, Portnov B.A.(2014). Evaluating the effect of vehicle impoundment policy on illegal construction and demolition waste dumping: Israel as a case study[J]. Waste Management, Vo.34, No. 8: pp. 1436-1445.

Taylor S, Todd P. (1995). An integrated model of waste management behavior A test of household recycling and composting intentions[J]. Environment and Behavior, Vo. 27, No. 5: pp. 603-630.

Tam W.Y.V, Li J.R and Cai H. (2014). System dynamic modeling on construction waste management in Shenzhen, China[J]. Waste Management & Research. Vo. 32, No. 5: pp. 441–453.

Tasaki T, Kawahata T, Osako M, Matsui Y, Takagishi S, Morita A, and Akishima S. (2007). A GIS-based zoning of illegal dumping potential for efficient surveillance[J].Waste Management, Vo. 27, No.2: pp. 256-267.

Witten I.H, Frank E.(2005). Data Mining: Practical Machine Learning Tools and Techniques with Java Implementation[M]. Morgan Kaufmann Publishers.

About the Authors

Bo Fan is a full Professor in School of International and Public Affairs, Shanghai Jiao Tong University. He has received PhD degree from Harbin Institute of Technology, and his major is Information Management and Public Administration. His research interests include E-government and emergency management. He has published in Decision Support Systems, Expert Systems with Applications, Natural Hazards, Information Development, Information Systems and E-business Management, Computer Aided Design and so on.

Long Chen is a master student in School of International and Public Affairs, Shanghai Jiao Tong University. His research interests include Environmental policy and Emergency management. He has published in many Chinese journals.

Yih Tng Chong is a research fellow in the Department of Industrial and Systems Engineering, Faculty of Engineering, National University of Singapore. He received his PhD from the Nanyang Technological University with his research in product design and development. His research interests include design and sustainability, with applications of computational intelligence methodologies. He has published in such journals as Research in Engineering Design, Advanced Engineering Informatics, Expert Systems with Applications and Computers & Industrial Engineering.

Zhou He is a research fellow in the Department of Industrial and Systems Engineering, Faculty of Engineering, National University of Singapore. He received his PhD from the Academy of Mathematics and Systems Science, Chinese Academy of Sciences. His research interests include management information systems, with applications of computational intelligence methodologies. He has published in such journals as European Journal of operational research and so on.