International Journal of Lean Thinking Volume 1, Issue 1 (June 2010)

Lean Thinking

journal homepage: www.thinkinglean.com/ijlt

Application of Optimisation-Based Data Mining Techniques To Tobacco Control Dataset

Z. Dzalilov *
Centre for Informatics and Applied Optimization, University of Ballarat, Ballarat, Victoria, Australia
z.dzalilov@ballarat.edu.au

J. Zhang
KOTCH Enterprises PTY. LTD, Victoria, Australia

A.M. Bagirov
Centre for Informatics and Applied Optimization, University of Ballarat, Ballarat, Victoria, Australia

M. Mammadov
Centre for Informatics and Applied Optimization, University of Ballarat, Ballarat, Victoria, Australia

ABSTRACT

Tobacco smoking is one of the leading causes of death around the world. Consequently, control of tobacco use is an important global public health issue. Tobacco control may be aided by development of theoretical and methodological frameworks for describing and understanding complex tobacco control systems. Linear regression and logistic regression are currently very popular statistical techniques for modeling and analyzing complex data in tobacco control systems. However, in tobacco markets, numerous interrelated factors nontrivially interact with tobacco control policies, such that policies and control outcomes are nonlinearly related. The use of linear and logistic regression is the refore fundamentally limited due to their inability to deal with these complex relationships. In this paper we aim to describe these relationships more effectively, by using global optimization-based approaches. We evaluate two distinct methods: the modified linear least square fit and a heuristic algorithm for feature selection based on optimization techniques. All these methods explore the relationship between features and classes, with the aim of determining contribution of specific features to the effectiveness of outcome. These methods allow consideration of datasets with an arbitrary number of classes. Our preliminary results indicate a possibility for a global optimal approach to covering all possible solutions in a complex tobacco control system. Compared with traditional statistical techniques, optimization-based methods therefore have the potential to be more effective analysis tools of complex tobacco control systems.

KEYWORDS

Tobacco control,
Data mining,
Global optimization,
Feature selection,
Linear least square fitting.

ARTICLE INFO

Received 21 April 2010
Received in revised form 20 June 2010
Accepted 22 June 2010
Available online 26 June 2010

1. Introduction

Control of cigarette smoking is an important global public health issue. The Framework Convention on Tobacco Control (FCTC) is the first international treaty devoted to public health and has propelled tobacco control into a new era. This framework was established in 2003 and has been ratified by up to 142 parties, representing 95% of the world's population (Zhang et al., 2007). Many countries, especially developed countries, have incorporated FCTC policies and recommendations into their laws.

*Corresponding Author

These countries have attempted to influence the behaviour of smokers by regulating and implementing diverse tobacco control policies, such as warning labels on cigarette packs, increases in the price of cigarettes and anti-smoking advertisements (David, 2000).

Controlling tobacco smoking and determining effective policies is difficult because of the complexity of human nature and behaviours. Nevertheless, there have been numerous attempts to describe and understand the effectiveness of tobacco control policies to smokers' quitting behaviour. Most of these attempts are based on traditional statistical techniques such as linear regression and logistic regression (Borland et al., 2009; Hammond et al., 2007). However, these techniques assume a linear sequence between research evidence outlining the need for action, the optimal forms of action, policy development, and the subsequent effectiveness of policies. Interestingly, even simple analyses have already demonstrated that behavioural changes of smokers in response to policy initiatives are often non-linear (Hammond et al., 2007). For instance, significant behavioural changes may be a consequence of what appear to be marginal increments in social/health policy. Also, the success of tobacco control is not the result of single policies, but is the outcome of interactions among various policies in various domains. Therefore, regression-based techniques are clearly inadequate for dealing with causalities among a set of such variables.

The International Tobacco Control Policy Evaluation Survey (ITC survey, 2010) is a recent coordinated international research and evaluation effort. This project provides massive survey data collected from many countries including Australia, for studying and evaluating the psychosocial and behavioural impact of diverse tobacco control policies to smokers across these countries (Figure 1).

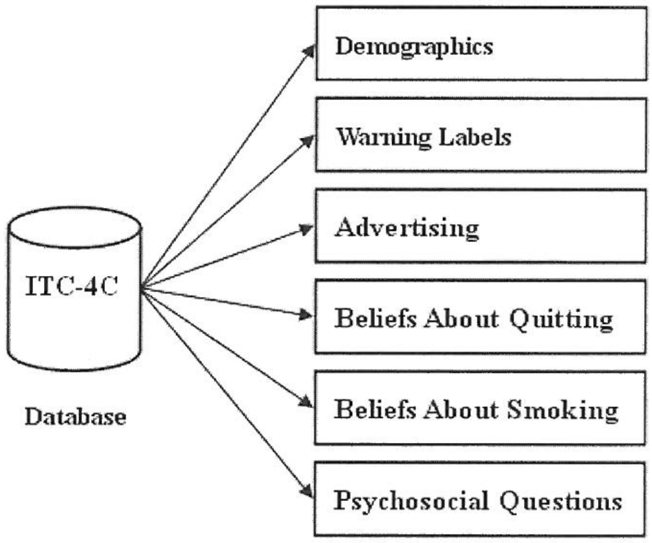

Figure 1. A structure of the data set under consideration

All respondents in each wave of each year are selected randomly from the population of individual countries above using random-digit-dialling methods. All aspects of the study protocol and survey measures are standardised across these countries. Many results on how diverse policy instruments have influenced smokers' quitting behaviours in the context of ITC project have been reported. However most of these results are based on traditional statistical techniques, such as linear regressions and logistic regressions. Although these traditional statistical models are very useful in monitoring and evaluating the impact of single policy instruments, they have fundamental limitations due to their inability to account for non-linear data relationships that may characterise many aspects of the tobacco domain.

In this paper, we propose new, optimization-based approaches for evaluating the effectiveness of diverse tobacco control policies on influencing smokers' quitting behaviours. The aim of this paper is to find a combination of smokers' responses to most significant questions about their demographics, anti-smoking advertisement, warning labels, and beliefs about quitting in the role of predicting the rate of quitting attempt and through such a study we can understand better the psychosocial and behavioural impact of diverse tobacco control policies to smokers. Our methods aim to answer a key question: "how can we predict the response of smokers to tobacco control policies?" Our results show that smokers' motivations and beliefs about quitting are the key factors for quitting attempt.

2. Methods and Results

We illustrate our techniques based on four sample data sets selected from five waves (2003-2007) of a cohort survey in Australia provided by the ITC project. The four sample data are Waves 2-3: n = 1458; Waves 3-4: n = 1477; Waves 4-5: n = 1260; and Waves 5-6: n = 1350. Here n is the number of reports in a data set. Australian smokers' responses to 75, 72, 64, 65 questions at 2, 3, 4, 5 waves were used to predict the rate of attempting to quit at the 3, 4, 5, 6 waves. The 75, 72, 64, 65 questions at 2, 3, 4, 5 waves covered many important aspects about smokers, including demographics, responses to warning labels, advertising, beliefs about quitting, beliefs about smoking and psychosocial questions. For example, among the 75 questions asked at wave 2, there are 4 questions about smokers' smoking status, heaviness of smoking, intention to quit and how hard is the whole day without smoking; there are 6 questions about harm of smoking; 12 questions about the influence of warning labels to smokers' quitting behaviours; 8 questions about responses to anti-smoking information; 4 questions about the confidence of quitting smoking; 15 questions about the reasons for trying to quit smoking; 6 questions about the enjoyment of smoking; 7 questions about the responses to public attitude and restriction and so on. While for the 72, 64, 65 questions at 3, 4, 5 waves, majority of the questions above are covered, the important aspects about smokers are almost the same as that asked at wave 2.

In order to analyse this data set we apply the following to optimization based algorithms:

- Linear least square fit.
- A heuristic algorithm for feature selection based on optimization techniques.

Next we give a brief description of these algorithms.

2.1 The Linear Least Squares Fitting (LLSF) (Yang, 1999; Yang and Liu, 1999).

Let M be the number of all features and C be the number of classes. Data is given in the form of two matrices:

Matrix $A=(a_{ij})$, $i=1,\ldots,N$, $j=1,\ldots,M$, where N is the number of samples.

Matrix $B=(b_{ik})$, $i=1,\ldots,N$, $k=1,\ldots,C$, where vector $(b_{i1},\ldots,b_{iC})$ describes class information for the row/sample i; $b_{ik}=1$ if sample i belongs to class k and $b_{ik}=0$ otherwise.

Consider matrix $X=(x_{jk})$, $j=1,\ldots,M$, $k=1,\ldots,C$, that describes the relationships between features and classes.

LLSF aims to find matrix X by minimizing the function $f(X)=\|AX-B\|^2$.

2.2 Importance of features by LLSF

Take any feature p and eliminate it from the list of features.

Denote $A(p)=(a_{ij})$, $i=1,\ldots,N, j=1,\ldots,p-1,p+1,\ldots,M$ and $X(p)=(x_{jk})$, $j=1,\ldots,p-1,p+1,\ldots,M,k=1,\ldots,C$.

Let $X^*(p) = \arg\min\{\|AX(p)-B\|^2: X(p)\in R^{M\times C}\}$. Matrix $X^*(p)$ can be used to predict all samples $i=1,\ldots,N$ using all features except p: $j=1,\ldots,p-1,p+1,\ldots,M$. Denote the average accuracy obtained in this way by $E(p)$. Clearly, the inequality $E(p_1) < E(p_2)$ for some features p_1 and p_2 means that the accuracy decreases more if we eliminate feature p_1 rather than p_2. Therefore, we can say that feature p_1 is more important than p_2; we write in this case $p_1 \succ p_2$. Arranging all features in a way that

$E(j_1)\leq E(j_2)\leq \ldots.. \leq E(j_M)$

We obtain the order of features by their importance in ascending order

$j_1 \succ j_2 \succ \ldots. \succ j_M$.

The illustration of the results of the applications of LLSF to the data set containing 1458 subjects with 71 features is given in Figure 2. Previously we applied LLSF to many different data sets, in particular to the data base on Cystic Fibrosis (see Hafen et al., 2008). Our preliminary results show that the methods applied are helpful in developing a clinical scoring system on Cystic Fibrosis.

3. A heuristic algorithm for feature selection based on optimization techniques

The feature selection problem involves the selection of a subset of features that will be sufficient in making predictions. We suggest an algorithm for the solution of the feature selection problem based on techniques of convex programming. We consider feature selection in the context of the classification problem. The algorithm suggested allows one to consider data sets with an arbitrary number of classes (Bagirov, 1999; Bagirov et al., 2002; Bagirov et al., 2003).

The algorithm calculates a subset of most informative features and a smallest subset of features. The first subset provides the best description of a dataset whereas the second one provides the description which is very close to the best one. A subset of informative features is defined by using certain thresholds. The values of these thresholds depend on the objective of the task.

The purpose of a feature selection procedure is to find as small a set as possible of informative features of the object under consideration, which describes this object from a certain point of view. The following issues are very important for understanding the problem.

1. It is convenient to consider (and define) informative features in the framework of classification. In other words it is possible to understand whether a certain feature is informative for a given example if we compare this example with another one from a different class.
2. Our goal is to find a sufficiently small set of informative features and to remove as many superfluous features as possible. Note that this problem can have many different solutions.
3. It follows from the above that the set of informative features, which describe a given object, is a categorical attribute of this object. This is also a fuzzy attribute in a certain informal sense. It leads to the following heuristic conclusion: it is useless to apply very accurate methods in order to find this set. However if we use heuristic not very accurate methods we need to have experimental confirmation of the results obtained.

In order to confirm that the results obtained are correct, we will use the following strategy: we consider a particular method for the search of a subset of informative features. Then we apply an auxiliary method based on a completely different idea to confirm that the discovered set of features describes the given object. If the auxiliary method confirms that the obtained set of features gives a correct description of the object then the result of the main algorithm can be accepted. Otherwise, further investigation is required.

We now describe the approach that will be used in this paper for feature selection (in the framework of classification). Assume we have a finite set A of vectors in n-dimensional space, R^n and we wish to give a compact description of this set in terms of informative variables. This description can be given by means of a single point, which is called the centre of the set A. To define the centre, consider the deviation $d(x,A)$ of a point $x \in R^n$ from the set A. By definition $d(x,A)$ is the sum of distances from x to points $a^i \in A$:

$$d(x,A) = \sum_{a^i \in A} \| x - a^i \| \tag{1}$$

Here $\| \cdot \|$ is a norm in R^n. In the sequel we will consider $\| \cdot \| = \| \cdot \|_p \ (p = 1,2)$, where

$$\| x - a^j \| = \| x - a^i \|_p = \left(\sum_{l=1}^{n} | x_l - a_l^i |^p\right)^{1/p}.$$

Definition 1. The point x^0 is called the centre of the set A (with respect to a norm $\| \cdot \|$) if x^0 minimizes the deviation from A, that is, x^0 is a solution of the problem

$$\sum_{a^i \in A} \| x - a^i \| \to \min \text{ Subject to } x \in R^n. \tag{2}$$

As a rule, one point cannot give a sufficiently precise description of the entire set A. To obtain this description, we need to have more points, which can be considered as centres of clusters of the set A. However we accept the hypothesis that for the search for such a categorical and fuzzy attribute as a set of informative features it is enough to consider only one representative of a set, namely its centre. It is assumed that this representative possesses some properties of the set, which allow one to replace the set itself by its centre in the study of the problem under consideration. The results of numerical experiments, which were confirmed by another (different) approach, demonstrate that this hypothesis leads to good results for many databases.

Consider now two subsets of the n-dimensional space R^n. To underline the dimension we include n in the notation. Thus we denote these sets by $A^1(n)$ and $A^2(n)$. Let $x^i(n)$ be the centre of the set $A^i(n), i = 1,2$. The quality of the description of the sets $A^1(n)$ and $A^2(n)$ by their centres can be expressed by numbers $N_1(n)$ and $N_2(n)$, where $N_1(n)$ is the number of the points from $A^1(n)$ which are closer to centre $x^2(n)$ of the other set $A^2(n)$ than to the centre $x^1(n)$ of $A^1(n)$. $N_2(n)$ has a similar meaning. We shall consider a four-tuple $(A^1(n), A^2(n), x^1(n), x^2(n))$. The number of "bad" points $N(n) = N_1(n) + N_2(n)$ can be considered as a certain numerical characteristic of this four-tuple. (A point, belonging to one of the sets is "bad" if this point is closer to the centre of the other set.)

We wish to find the most informative features which allow us to distinguish sets $A^1(n)$ and $A^2(n)$ and remove the least informative features. Since $x^1(n)$ and $x^2(n)$ are representatives of the sets $A^1(n)$ and $A^2(n)$, respectively, which can replace the corresponding sets under consideration, we can assume that the closest coordinate of centres indicates the least informative feature. Then we can try to remove this coordinate. If we remove it, we will have new sets $A^1(n-1)$ and $A^2(n-1)$ and we can find centres $x^1(n-1)$ and $x^2(n-1)$ of these sets. Consider a new four-tuple $(A^1(n-1), A^2(n-1), x^1(n-1), x^2(n-1))$ and calculate the numerical characteristic $N(n-1) = N_1(n-1) + N_2(n-1)$ of this four-tuple. If this characteristic is close enough to $N(n)$, we can deduce that the eliminated coordinate has little influence on this characteristic. So this coordinate need not belong to a set of informative features. On the other hand if the difference between $N(n)$ and $N(n-1)$ is sufficiently large, then we can not remove even the closest coordinate of centres, so all n coordinates belong to a set of informative features.

To be more precise we need to define what "sufficiently large" means. To achieve this we will introduce certain thresholds. The procedure is based on an inner description of a set by means of its centre and on comparison of two sets by means of their centres. We can treat this approach as an inner approach to feature selection. The opposite to an inner approach is an outer approach which is based on a separation of two sets. The main idea behind the outer approach is the following. Having two sets A_1 and A_2 we can find a discrimination function c such that $c(a^1) > 0$ for $a^1 \in A_1$ and $c(a^2) < 0$ for $a^2 \in A_2$. Then if a new point is presented we can calculate $c(a)$. If $c(a) > 0 (c(a) < 0$, respectively) then we bring to $A_1 (A_2$, respectively). Various versions of this approach have been studied in (Abello et al., 2001) and very sophisticated methods for the determination of a discrimination function c have been proposed. However in some instances we can use very simple discrimination functions. One of them is applied in this paper, when we consider an auxiliary algorithm for confirmation of our main feature selection algorithm.

We consider a dataset which contains m classes, that is, we have m nonempty finite sets $A_i, i = 1,\dots,m$ in R^n consisting of $r_i, i = 1,\dots,m$ points, respectively. Using the idea, presented above we suggest an algorithm for the solution of the feature selection problem. Below we will consider two cases for $p = 1$ and $p = 2$. Let

$N_1 = \{j : j = 1,\dots,| A_1 |\},\ N_i = \{j : j = | A_{i-1} | +1,\dots,| A_{i-1} | + | A_i |\}, i = 2,\dots,m,$

where $| A_i |$ denotes the cardinality of the set $A_i, i = 1,\dots,m$. First we will consider the case when $m = 2$. Let $\varepsilon > 0$ be some tolerance and $T_i \in \{1,2,\dots\}, i = 1,2,3$ be the thresholds.

Algorithm 1. Feature selection

Step 1. Initialization. Set $k = 0, I_k = \{1,\dots,n\}$.

Step 2. Determination of centres of the sets $A_i, i = 1,2$. (See Definition 1). Compute the centre of A_i by solving the following problems of convex programming:

$$\sum_{j\in N_i} \| x^i - a^j \|_p \rightarrow \min \quad \text{Subject to } x^i \in R^n, p = 1,2. \tag{3}$$

Here $\| x^i - a^j \|_1 = \sum_{l\in I_k} | x_l^i - a_l^j |, \| x^i - a^j \|_2 = [\sum_{l\in I_k} (x_l^i - a_l^j)^2]^{1/2}$.

Step 3. Find points of the set $A_i, i = 1,2$, which are closer to the centre of the other set. Let $x^{i*}, i = 1,2$ be solutions to the problem (3). Compute the sets:

$N_1^k = \{j \in N_1 : \| x^{2*} - a^j \|_p \le \| x^{1*} - a^j \|_p\},\ N_2^k = \{j \in N_2 : \| x^{1*} - a^j \|_p \le \| x^{2*} - a^j \|_p\}$.

Set $N_3^k = N_1^k \cup N_2^k$. If $k = 0$ then go to Step 5, otherwise go to Step 4.

Step4. Calculate $L_i = \max\{| N_i^t | : t = 0,\dots,k\}, i = 1,2,$

$L_3 = \max\{| N_1^t | + | N_2^t | : t = 0,\dots,k\}$.

If $\max\{L_i - T_i : i = 1,2,3\} > 0$ then I_{k-1} is a subset of most informative features and the algorithm terminates. Otherwise go to Step 5.

Step 5. To determine the closest coordinates. Calculate $d_0 = \min\{| x^{1*}_l - x^{2*}_l | : l \in I_k\}$ and define the following set: $R_k = \{l \in I_k : | x^{1*}_l - x^{2*}_l | \le d_0 + \varepsilon\}$.

Step 6. Construct the set: $I_{k+1} = I_k \setminus R_k$. If $I_{k+1} = \varnothing$ then I_k is the subset of most informative features. If $| I_{k+1} | = 1$ then I_{k+1} is the subset of most informative features. Then the algorithm terminates otherwise set $k = k + 1$ and go to Step 2.

Remark 1. An algorithm for the case when the number of classes $m > 2$ can be obtained from the Algorithm 1 by replacing Steps 3, 4 and 5 by the Steps 3', 4' and 5', respectively.

Step 3'. To find points of a set $A_i, i = 1,\dots,m$, which are closer to the centres of other sets? Let $x^{i*}, i = 1,\dots,m$ be solutions to the problem (3). Compute the sets:

$N_i^k = \{j \in N_i : \min\{\| x^{t*} - a^j \|_p : t = 1,\dots,m, t \ne i\} \le \| x^{i*} - a^j \|_p\}, i = 1,\dots,m.$

Set $N_{m+1}^{\ k} = \cup\{N_i^{\ k} : i = 1,\dots,m\}$. If $k = 0$ then go to Step 5', otherwise go to Step 4'.

Step 4. Calculate $L_i = \max\{| N_i^{\ t} | : t = 0,\dots,k\}, i = 1,\dots,m,$

$$L_{m+1} = \max\{\sum_{i=1}^{m} | N_i^{\ t} | : t = 0,\dots,k\}.$$

If $\max\{L_i - T_i : i = 1,\dots,m+1\} > 0$ then I_{k-1} is a subset of most informative features and the algorithm terminates. Otherwise go to Step 5'.

Step 5'. To determine the closest coordinates. Calculate

$d_l = \max\{| x^{i^*}{}_l - x^{t^*}{}_l | : i,t = 1,\dots,m\}, \quad d_0 = \min\{d_l : l \in I_k\}$

and define the following set: $R_k = \{l \in I_k : d_l \le d_0 + \varepsilon\}$.

Remark 2. It should be noted that the subset of most informative features calculated by the Algorithm 1 depends on the vector of thresholds $T = (T_1, T_2, \dots, T_{m+1})$. In numerical experiments we shall consider two cases:

1. $T_i = | A_i | / 100, T_{m+1} = \sum_{i=1}^{m} T_i$; 2. $T_i = 2 | A_i | / 100, T_{m+1} = \sum_{i=1}^{m} T_i$.

We shall use some indicators in order to confirm the results obtained by Algorithm 1. One of them is the accuracy e_i for the set $A_i, i = 1,\dots,m$ and total accuracy e_{tot} for all sets. Assume that misclassified points, that is, points from a given set, which are closer to the centre of the other set, are known. Then the accuracy e_i for the set A_i is calculated as follows: $e_i = 100(| A_i | - m_i) / | A_i |$, (4)

where m_i is a number of misclassified points for the set $A_i, i = 1,\dots,m$. In the same manner we can define total accuracy e_{tot} for all sets:

$$e_{tot} = 100(| A \ | - M) / | A \ |, \tag{5}$$

where $A = \sum_{i=1}^{m} | A_i \ | \ , M = \sum_{i=1}^{m} m_i$.

As mentioned above, the centre of the set as a rule can not give a sufficiently precise description of the entire set, so the accuracies e_i and e_{tot} as a rule are not very high. However, we use this indicator in order to recognize the importance of the removed coordinate at each iteration of the algorithm. That is, we again include the dimension n of the set under consideration in the notation. Assume that after some iteration we have n-dimensional vectors. Denote the corresponding sets by $A_i(n)$ and let $e_i(n)$ be the accuracy for the set $A_i(n)$. Assume that a particular coordinate was removed, and then we have the set $A_i(n-1)$ with the accuracy $e_i(n-1)$. The approximate equality $e_i(n) \cong e_i(n-1)$ can be considered as confirmation that the removed coordinate does not have any influence on the structure of the set A_i. On the contrary, if the difference between $e_i(n)$ and $e_i(n-1)$ sharply increases then we can suppose that the structure of sets $A_i(n)$ and $A_i(n-1)$ is different, so the coordinate that was removed is indeed informative.

We can consider different indicators, which confirm the outcome of Algorithm 1. One of them is based on a very simple and rough classification algorithm. This algorithm provides a classification with not very high accuracy, which is compatible with the accuracy of the description of a set by its centre. However this algorithm does not require very much computational time, so it can be used to verify the results obtained by the feature selection algorithm. We suppose that the data set under consideration contains two classes $A_i, i = 1,2$. We define the following quantity for any $x \in R^n$:

$$\omega(x) = \frac{\rho(x, A_1)}{1+\rho(x, A_2)} - \frac{\rho(x, A_2)}{1+\rho(x, A_1)}.$$

where $\rho(x, A_i) = \min\{\| x-y \|: y \in A_i\}, i = 1,2.$ It is easy to see that $\omega(x) = -\rho(x, A_2) < 0$ for all $x \in A_1$ and $\omega(x) = -\rho(x, A_1) < 0$ for all $x \in A_2$. Then knowing training sets A_1 and A_2 we can suggest the following algorithm for classification. Let B be a set of new observations and $N =| B |$.

Algorithm 2. Classification algorithm

Step 1. Initialization. Set $k = 1$.

Step 2. Take $x^k \in B$ and calculate $\rho_k = \rho(x^k, A_i) = \min\{\| x^k - a \|: a \in A_i\}, i = 1,2.$

Step 3. If $\frac{\rho_1}{1+\rho_2} < \frac{\rho_2}{1+\rho_1}$ then $x^k \in A_1$, otherwise $x^k \in A_2$. Set $k = k+1$. If $k \le N$ then go to Step 2, otherwise stop.

4. Numerical Results

As a preliminary work, we applied both algorithms to data sets containing:

- 1458 subjects, with 71 features
- 1477 subjects, with 69 features
- 1260 subjects, with 60 features
- 1350 subjects, with 60 features

Results obtained by LLSF algorithm are illustrated in Figure 2. This figure shows dependence of the classification accuracy on the number of features. One can see that this algorithm does not allow one to find the subset of most informative features.

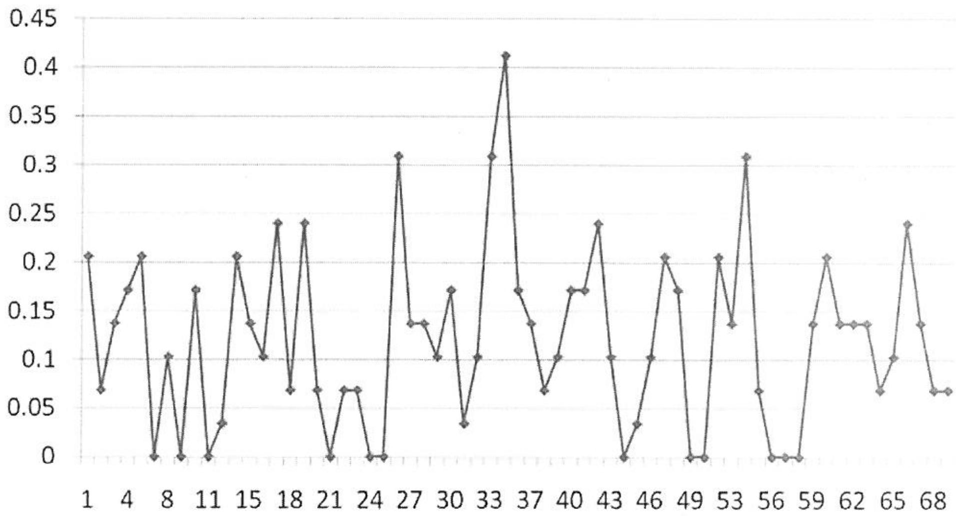

Figure 2. Wave 2-3

Illustrations of numerical results of applications of the feature selection algorithm to the different waves of the data set are given on the Figures 3-6, below. These figures show dependence of the classification accuracy on the number of features.

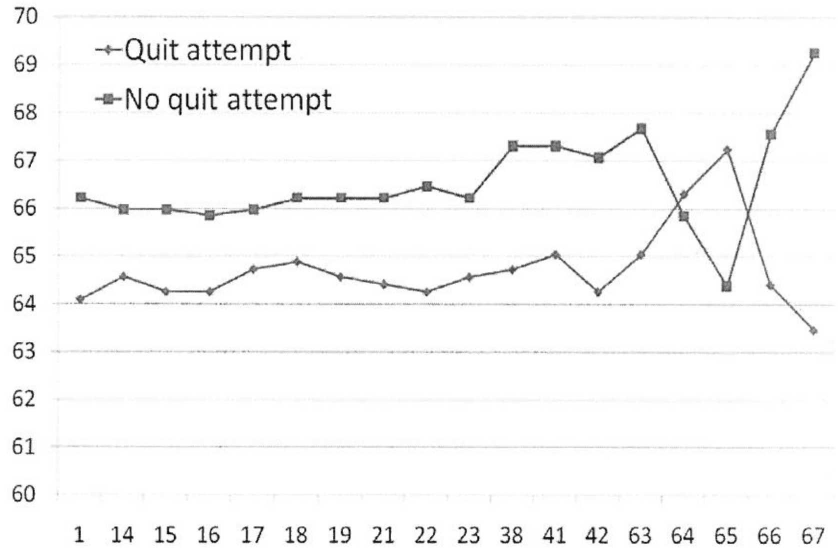

Figure 3. Wave 2-3

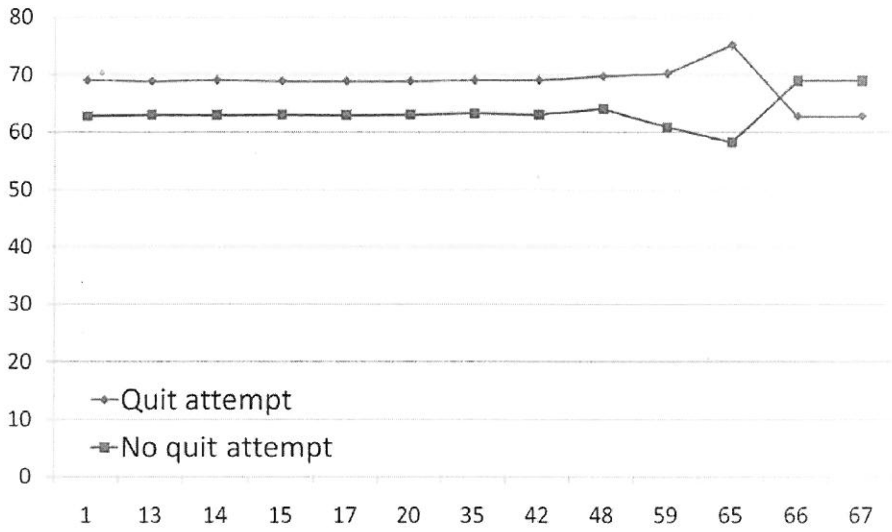

Figure 4. Wave 3-4

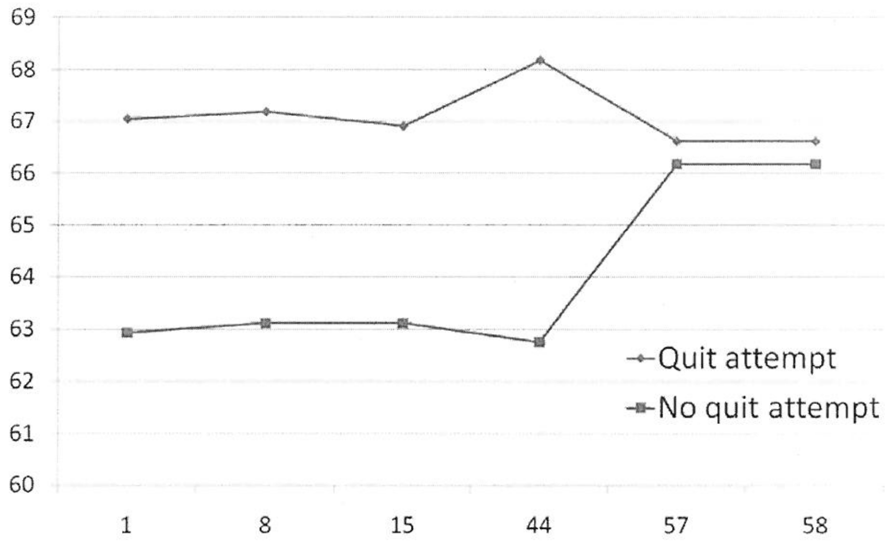

Figure 5. Wave 4-5

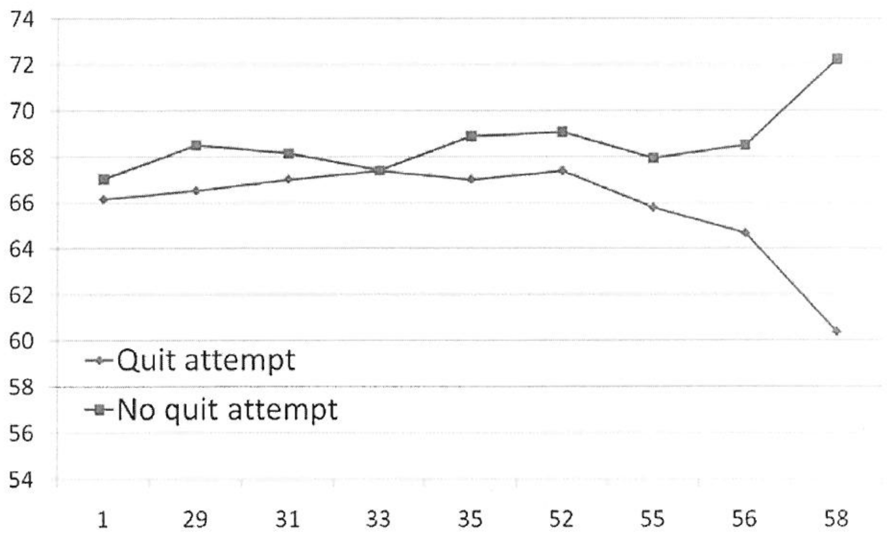

Figure 6. Wave 5-6

Some examples: the meanings of the significant features for quit attempt

- 5 Thoughts about danger of smoking;
- 7 Thoughts about harm to self;
- 35 Confidence to quit smoking;
- 37 Perception of quitting difficulty;
- 66 Disapproval of smoking in society.

Results presented in Figures 3-6 allow us to draw the following conclusions:

- Heuristic algorithm revealed complementary pattern of features for quitters and non-quitters.
- Overlapping features are likely to be important in tobacco control programs.

5. Discussions and conclusions

As a result of the feature selection algorithm we found a number of significant features that are associated with quit attempts. The features in the neighbourhood of the maximal point can also be considered to be associated with quit attempts, with a reasonable level of accuracy. The maximum accuracy is 67.24. The list of smokers' response to most significant questions at waves 2,3,4,5 in predicting the rate of quitting attempt at waves 3,4,5,6 are "think about danger of smoking, think about harm of smoking, sure you will success at quitting, how hard to completely quit, society disapprove"; "sure you will success at quitting, how hard to complete quit, how easy or hard to stop quitting permanently"; "sure you will success at quitting, how hard to complete quit, how easy or hard to stop quitting permanently"; and "sure you will success at quitting, how hard to complete quit" respectively. These findings from all waves survey data have demonstrated that among many questions we asked at each wave surveys, the most significant features for pushing smokers to make a quit attempt are focused on the following aspects: knowledge about the harm of smoking, worry about health, confidence about quitting, addiction to smoking". Our findings are consistent with that we have found before by using other methods. Our methods have answered that smokers' motivations, knowledge about harm of smoking, beliefs about quitting and so on are key factors for making a quitting attempt. If we consider 64 to be sufficiently accurate, then the number of significant features increases. We should mention as well that from the 2 methods above, the feature selection algorithm appears to perform best.

Compared with the traditional statistical techniques, the new methods have potential to become a good theoretical and methodological framework for modelling and analysing complex tobacco control systems. We have obtained some promising preliminary results for covering many important solutions in tobacco control. The results of analysis of the given data set are most likely to develop new models for a new survey, more accurate than the previous one.

Acknowledgements

We thank the ITC Project for permission to use data from the ITC Four Country Survey for the analysis of this paper. Data collection for the ITC Four Country Survey has been supported in part by grants from the National Cancer Institute of the United States (R01 CA 100362), the Roswell Park Transdisciplinary Tobacco Use Center (P50 CA111236), and (P01 CA138389; Roswell Park Cancer Institute, Buffalo, New York); and by the Medical Research Council of Australia (265903 and 450110). We also would like to thank Prof. Mary Thomson, Department of Statistics and Actuarial Science, University of Waterloo,Waterloo, Ontario CANADA, and Dr. Ron Borland, Cancer Council Research Institute, Melbourne, Australia for their continuos support.

References

Abello J., Pardalos P.M., Resende M.G.C. Handbook of Massive Datasets. Kluwer Academic Publishers: Dordrecht, 2001.

Bagirov A.M., Rubinov A.M., Yearwood J. A Global Optimisation Approach to Classification. Optimization and Engineering 2002; Vol. 3, No. 2; 129-155.

Bagirov A.M., Rubinov A.M., Yearwood J. A heuristic algorithm for feature selection based on optimisation techniques, In: Heuristic and Optimization for Knowledge Discovery, C. Newton, H. Abbas and R. Sarker (eds.), Idea Group Publishing, 2002; 13-26.

Bagirov A. M., Rubinov A. M., Soukhoroukova N.V. and Yearwood J, Supervised and unsupervised data classification via nonsmooth and global optimisation, TOP: Spanish Operations Research Journal 2003; Vol. 11, No. 1; 1-93.

Borland R., Yong H. H., Geoffrey N. W., Fong G. T., Hammond D., Cummings K.M., Hosking W., McNeill A. How reaction to cigarette packet health warnings influence quitting: findings from the ITC four country survey. 2009; preprint.

David R. M. 50 years of reporting on tobacco and health. British Medical Journal 2000; 320; 74.

Hafen G., Hurst C., Yearwood J., Mammadov M., Smith J., Dzalilov Z., Robinson P. A new clinical scoring system in Cystic Fibrosis: Statistical tools for database analysis. Medical Informatics and Decision Making 2008; 8; 44; http://www.biomedcentral.com/1472-6947/8/44

Hammond D., Fong G. T., Borland R., Cummings K. M., McNeill A., Driezen P. Text and graphic warnings on cigarette packages: findings from the international tobacco control four country studies. American Journal of Preventive Medicine 2007; 32; 202-209.

ITC International tobacco control policy evaluation survey. 2002-2010; http://www.itcproject.org.

Mammadov M. A., Rubinov A. M., Yearwood J. The study of drug-reaction relationships using global optimization techniques. Optimization Methods and Software 2007.

Yang Y. An Evaluation of Statistical Approaches to Text Categorization. Information Retrieval 1999; 1; 69-90.

Yang Y., Liu X. A re-examination of text categorization methods. Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval Berkeley, California; 1999; 42-49.

Zhang J. Y., Young D., Coghill K., Petrovic-Lazarevic S., Borland R., Yeh C. H., Bedingfield S. A New Theoretical Framework for Modelling and Analysing Complex Tobacco Control Systems. Achieving Competitive Advantage Through Managing Global Resources, Readings Book of The Global Business and Technology Association's Ninth Annual International Conference (GBATA 2007) Taipei, Taiwan; 2007; 811-817.