Artificial Intelligence in Medicine (2007) **41**, 251–262

ELSEVIER

ARTIFICIAL INTELLIGENCE IN MEDICINE

http://www.intl.elsevierhealth.com/journals/aiim

Feature selection and classification model construction on type 2 diabetic patients' data

Yue Huang [a,*], **Paul McCullagh** [b], **Norman Black** [b], **Roy Harper** [c]

[a] *Department of Computing, Faculty of Engineering, Imperial College London, South Kensington, London SW7 2AZ, UK*
[b] *School of Computing and Mathematics, Faculty of Engineering, University of Ulster, Jordanstown BT37 0QB, UK*
[c] *The Ulster Hospital, Dundonald, Belfast BT16 0RH, UK*

Received 2 November 2006; received in revised form 19 June 2007; accepted 6 July 2007

KEYWORDS
Type 2 diabetes;
Blood glucose;
Data mining;
Classification;
Feature selection

Summary

Objective: Diabetes affects between 2% and 4% of the global population (up to 10% in the over 65 age group), and its avoidance and effective treatment are undoubtedly crucial public health and health economics issues in the 21st century. The aim of this research was to identify significant factors influencing diabetes control, by applying feature selection to a working patient management system to assist with ranking, classification and knowledge discovery. The classification models can be used to determine individuals in the population with poor diabetes control status based on physiological and examination factors.
Methods: The diabetic patients' information was collected by Ulster Community and Hospitals Trust (UCHT) from year 2000 to 2004 as part of clinical management. In order to discover key predictors and latent knowledge, data mining techniques were applied. To improve computational efficiency, a feature selection technique, feature selection via supervised model construction (FSSMC), an optimisation of ReliefF, was used to rank the important attributes affecting diabetic control. After selecting suitable features, three complementary classification techniques (Naïve Bayes, IB1 and C4.5) were applied to the data to predict how well the patients' condition was controlled.
Results: FSSMC identified patients' 'age', 'diagnosis duration', the need for 'insulin treatment', 'random blood glucose' measurement and 'diet treatment' as the most important factors influencing blood glucose control. Using the reduced features, a best predictive accuracy of 95% and sensitivity of 98% was achieved. The influence of factors, such as 'type of care' delivered, the use of 'home monitoring', and the importance of 'smoking' on outcome can contribute to domain knowledge in diabetes control.

* Corresponding author. Tel.: +44 20 75948382; fax: +44 20 75818024.
E-mail address: y.huang@imperial.ac.uk (Y. Huang).

0933-3657/$ – see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.artmed.2007.07.002

Conclusion: In the care of patients with diabetes, the more important factors identified: patients' 'age', 'diagnosis duration' and 'family history', are beyond the control of physicians. Treatment methods such as 'insulin', 'diet' and 'tablets' (a variety of oral medicines) may be controlled. However lifestyle indicators such as 'body mass index' and 'smoking status' are also important and may be controlled by the patient. This further underlines the need for public health education to aid awareness and prevention. More subtle data interactions need to be better understood and data mining can contribute to the clinical evidence base. The research confirms and to a lesser extent challenges current thinking. Whilst fully appreciating the requirement for clinical verification and interpretation, this work supports the use of data mining as an exploratory tool, particularly as the domain is suffering from a data explosion due to enhanced monitoring and the (potential) storage of this data in the electronic health record. FSSMC has proved a useful feature estimator for large data sets, where processing efficiency is an important factor.
© 2007 Elsevier B.V. All rights reserved.

1. Introduction

Diabetes is the most common endocrine disease in all populations and all age groups. According to the World Health Organisation, it affects around 194 million people worldwide, and that number is expected to increase to at least 300 million by 2025. Diabetes has become the fourth leading cause of death in developed countries and there is substantial evidence that it is reaching epidemic proportions in many developing and newly industrialized nations [1], with evidence pointing to avoidable factors such as sedentary lifestyle and poor diet.

Diabetes describes a metabolic disorder characterized by chronic hyper-glycaemia with disturbances of carbohydrate, fat and protein metabolism resulting from defects in insulin secretion, insulin action, or both. The long-term complications include progressive development of the specific complications of retinopathy with potential blindness, nephropathy that may lead to renal failure, and/or neuropathy with risk of foot ulcers, amputation, Charcot joints, and features of autonomic dysfunction, including sexual dysfunction, known as micro-vascular complications. People with diabetes are also at a greatly increased risk of cardiovascular, peripheral vascular, and cerebro-vascular disease [2], known as macro-vascular complications.

There are two main classes of diabetes, which are diagnosed ultimately by the severity of the insulin deficiency. Insulin-dependent diabetes mellitus or type 1 diabetes is an insulinopenic state, usually seen in young people, but it can occur at any age [3]. Non-insulin-dependent diabetes mellitus or type 2 diabetes is the more common metabolic disorder that usually develops in overweight, older adults, but an increasing number of cases occur in younger age groups. Pinhas-Hamiel and Zeitler [4] predict serious public health challenges given the rise in paediatric cases and the poor medication adherence in teens. The prevalence of type 2 diabetes has risen from 3% to 45% of adolescent diabetes in the last 15 years. In this age group for girls, complications in pregnancy add to the social cost. In Northern Ireland in 2004/5, 25% girls and 20% boys in the age range 4.5—5.5 years were classified as overweight or obese and research indicates that 85% of obese children will become obese adults.

In type 2 diabetes, the pancreas may produce adequate amounts of insulin to metabolize glucose (sugar), but the body is unable to utilize it efficiently. Over time, insulin production decreases and blood glucose levels rise. Patients with type 2 diabetes do not require insulin treatment to remain alive, although up to 20% are treated with insulin to control blood glucose levels. Type 2 diabetes accounts for up to 85% of the diabetic population in most countries; it probably affects 5—7% of western populations, 10% of people over 65 years of age, and up to 50% of the cases may be currently undiagnosed [5]. The peak age of onset of type 2 diabetes is 60 years old; most subjects are diagnosed after 40 years of age.

The burden of complications can be considerable both for the individual concerned and the health service in general. In the United Kingdom, economic costs are estimated at £2.5 billion annually. Many aspects of these complications can be limited, even prevented in some instances, with good early management of the condition, in particular the effective control of blood glucose levels. Computer-based tools can assist healthcare professionals to better manage people with type 2 diabetes, in order to reduce complications and improve quality of life.

Better control of blood glucose reduces the risk of diabetic-related complications significantly. So understanding the major factors that determine overall control is important for clinicians in the prevention of diabetic complications, and will assist the patient with self management [6—9]. The

attending physician plays an important role in providing information to reduce those risk factors. It is up to the physician to warn patients at risk about the major causes of a particular blood glucose control status and the degree of risk that they are facing. In an attempt to achieve more effective diabetes management, we have applied data mining techniques to a 'working' database to verify the important risk factors and obtain information that may be unknown to the clinicians. The database contains 'noisy' clinical data and provides a stringent test for the application of data mining in clinical practice.

2. Data mining in diabetes

In modern medicine, large amounts of data are generated, but there is a widening gap between data collection and data comprehension. It is often impossible to process all of the data available and to make a rational decision on basic trends. Thus, there is a growing pressure for intelligent data analysis such as data mining to facilitate the creation of knowledge to support clinicians in making decisions.

Data mining techniques, e.g. feature ranking and classification model construction could be used in such databases to support other research studies. For example, Nissen and Wolski [10] have indicated that rosiglitazone, an oral hypoglycaemic drug, increases cardiovascular risk compared to other therapies or placebo, by providing a meta-analysis of treatment trials. Their analysis was at trial-level rather than patient level. There was no standard method for validating outcomes across trials and the total number of events was relatively small. If a classification model was able to identify such risks in a clinical database, it would clearly enhance its efficacy as a practical tool, and add to the evidence.

The role of data mining is to extract interesting (non-trivial, implicit, previously unknown and potentially useful) patterns or knowledge from large amounts of data, in such a way that they can be put to use in areas such as decision support, prediction and estimation. The data is often voluminous but may be of low value as no direct use can be made of it; indeed in some cases, it may be information hidden in the data that is useful in extending knowledge.

2.1. Classification

The purpose of classification is to predict categorical class labels based on the classification model built by training data. The aim of this study is to distinguish patients with bad blood glucose control from those with good blood glucose control based on physiological and examination data. The intention is to improve the quality of treatment by providing support to the expert. Feature selection can identify the most useful information from the data, and reduce the dimensionality in such a way that the most significant aspects of the data are represented by the selected features [11].

Classification algorithms can be used to compare the effects of feature selection with no feature selection. Three classification methods were used in this research: a probabilistic learner, Naïve Bayes [12], a decision tree learner, C4.5 [13] and an instance-based learner, IB1 [14]. These algorithms have proved effective in practice [15—19] and in particular in the clinical domain [20—23]. Indeed it would be possible to use other classification techniques such as support vector machines [68], neural networks [13,36] and rough sets [13]. Su et al. [24] used four data mining approaches (neural network, decision tree, logistic regression and rough sets) to select the relevant features for the diabetes diagnosis, and also evaluated their performance. For example, the features selected by the neural network were evaluated by neural network; that is to say, every method is both a feature selector and a classifier. In our research FSSMC was used as a general feature selector, and tested using three classifiers to assess its performance. The selected factors have been also been verified by the diabetic expert.

Every technique has its specific advantages and disadvantages, and is applicable for different research problems. Naïve Bayes, IB1 and C4.5 have been selected because they have long standing tradition in classification studies. The algorithms are straightforward to implement, producing relatively fast code [22]. C4.5 has an inherent feature selection mechanism, which allows the investigation of feature selection techniques for different types of classifier. It induces knowledge that is easy to interpret by the expert. IB1 and Naïve Bayes have proved popular because they have been shown to achieve good classification performance [20].

2.2. Feature selection

To improve the efficiency of classification algorithms, feature selection is used to identify and remove as much of the irrelevant and redundant information as possible. In the treatment of diabetes, hundreds of attributes are routinely collected but only a small number are used, i.e. the clinicians routinely perform ad-hoc feature selection. Being a real world problem, a large number of noisy, irrelevant and redundant features are in the data. All the features in the database were not specified for blood glucose control

prediction and much irrelevant information has been collected. Clearly irrelevant attributes such as 'driving status' have been removed and other clinical attributes with the help of the diabetic expert where necessary, leaving 47 features to be ranked, out of 410 which were stored for clinical management.

FSSMC [25] an optimisation of ReliefF [26], which has been successfully applied in data mining applications [27–29], was used to investigate those important factors in the type 2 diabetes data set. Two important issues remain problematic in ReliefF:

- ReliefF did not address the problem of multi-valued attributes [26]. At present, the similarity measurement applied in ReliefF is a numerical method, and if the two selected instances have the same categorical value, the result of difference function is 0, otherwise is 1. This definition cannot measure the contribution of multi-class ($\geq$3) values to class labels [30,31], and will underestimate numerical attributes [29].
- The setting of the number of instances m, sampled from the data set, which defines the number of iterations. There is a trade off between the use of more instances and the efficiency of computation. Two approaches have been applied conventionally: either m is set empirically or the entire data are analyzed. With m being a constant, the time complexity for a data set with n instances and a attributes becomes $O(n{\cdot}a)$. For large n, it often requires that $m \ll n$ for high efficiency. FSSMC finds 'typical' instances that can represent the whole data set, so that the performance can be close to the method using n instances. The parameter m (sampled instances) is generated automatically and the computational efficiency of the algorithm is improved.

A frequency based encoding scheme [31] has been used for data transformation to solve the problem of handling categorical data. Its scalability has been optimized by a distance-based re-sampling method which is inspired by active learning techniques [15,32–33]. The advantage is that active learning avoids pure random sampling used in traditional data mining approaches and is realized by selective sampling [34], as instances are not uniformly distributed and some instances are more representative than others [35]. If one can identify and select representative instances, fewer instances are needed to achieve similar performance. FSSMC maintains the classification accuracy (CA) of ReliefF, while improving its computational efficiency, on large databases.

3. Data description

The data were not specifically collected for a research study. As part of routine patient management, UCHT collected diabetic patients' information from 2000 to 2004 in a clinical information system (Diamond, Hicom Technology). The data contained physiological and laboratory information for 3857 patients, described by 410 features. The patients included not only type 2 diabetic patients, but also type 1 and other types of diabetes such as gestational diabetes.

It is very important to examine the data thoroughly before undertaking any further steps in formal analysis. Distorted data, incorrect choice of steps in methodology, misapplication of data mining tools, too idealized a model, a model which goes beyond the various sources of uncertainty and ambiguity in the data all represent possibilities for taking the wrong direction in a data mining process. Therefore, data mining is not just a matter of simply applying a directory of tools to a given problem, but rather a process of critical assessments, exploration, testing and evaluation. The data should be well defined, consistent and non-volatile in nature [36]. For this reason, preprocessing consisting of data integration, transformation and reduction, was applied to the data.

The final database contained 2064 type 2 diabetic patients' information; 1148 males and 916 females. The patients ranged in age from 20 to 96 years old, and average age was 67 years (47 patients are younger than 40 years). The standard used to define a label as 'good' and 'bad' blood glucose control was based on the corresponding laboratory HbA1c test value of an individual. This is a measure of glycated haemoglobin used to identify the plasma glucose concentration over a 3–4-month period of time, and accepted as the best indicator of control. The method was to compare the patient's HbA1c test value and his/her target HbA1c value (set by the clinicians based on expertise and judgment); if the laboratory test value was higher than the target, the patient was partitioned to the 'bad' control group; vice versa. Generally, the target HbA1c value was 7%, but varied according to individual condition. The selection of the cut-point was defined based on clinical outcomes and complication rates in type 2 diabetes. The 410 features in the database include patient characteristics, treatment, complication care, and physical and laboratory findings. Obvious irrelevant and sparse (more that 50% missing data) features were discarded from further investigation. The features used in the experiments were recommended by the clinician and verified by guidelines for diagnosis and management. The resulting data set had an average 7.8% of missing values. Among the initial 47 features, there were 2 attributes

Table 1 Mean, standard deviation (S.D.) and percentage missing values of nine key predictors (recommended by the clinician)

Patients' features	Mean	S.D.	Missing (%)
Age (year)	66.77	10.47	0.00
Diagnosis duration (year)	9.45	7.35	7.55
Haemoglobin, HbA1c (%)	8.35	1.64	0.00
Random blood glucose, LabRBG (%)	11.68	5.48	4.31
Triglycerides (mmol/l)	2.85	2.11	5.21
Cholesterol (mmol/l)	5.09	1.28	6.75
BMI (kg/m^2)	31.21	8.69	13.14
Systolic blood pressure (mmHg)	145.54	22.22	5.98
Diastolic blood pressure (mmHg)	80.91	12.60	6.13

with 30—40% missing data, 3 with 20—30%, 5 with 10—20% and 13 with 1—10%, and 23 categorical attributes and 24 numerical attributes. There were 20,876 records in the data set, 22.8% good diabetes control status and 77.2% in bad status. Each patient had one or more records.

Table 1 provides summaries of mean and standard deviations of some important numerical attributes, and Table 2 highlights the statistical difference of these attributes between the good and bad control groups. Average population age is 63.1 $\pm$ 10.6 years for the good control group, and 67.9 $\pm$ 10.2 years for the bad control group. Table 2 illustrates that those cases in the good control group exhibit younger age, shorter diagnosis duration, lower body-mass index (BMI), lower cholesterol and lower blood pressure levels compared to those in the bad control. Younger adults tend to manage their blood glucose level better; higher cholesterol and triglycerides level are associated with bad blood glucose control; a well-controlled BMI may help an individual manage his/her diabetes better.

4. Experimental methodology

Features were evaluated using FSSMC, which measures the usefulness of a feature by observing the relation between its value and the patient's outcome. Intuitively, if there is a group of patients with similar values, the observed feature is valuable as a predictor if it has different values on pairs of patients with different outcomes, but the same value on pairs with the same outcome. Features with a negative value may be considered to be irrelevant. Features with the highest score are presumed to be the most sensitive and contributing most to the outcome prediction [37].

There are many possible measures for evaluating feature selection algorithms and classification models [38]. CA was used to evaluate the performance of FSSMC. A classifier is expected to preserve the CA with the reduced set of features or to improve it due to the elimination of noisy and irrelevant features that may mislead the learning process. Sensitivity [39] measures the fraction of positive cases that are classified as positive. Specificity measures the fraction of negative cases classified as negative. For distinguishing patients with bad blood glucose control in the population, a high sensitivity is more important than specificity.

Because it is difficult to estimate the correct number of predictors [40], different feature subsets were selected for each of Naïve Bayes, IB1 and C4.5 to determine which set gave the best performance. Ten-fold cross-validation was used as the sampling

Table 2 Diabetic patients' features indicating good/bad blood glucose control

Patients' features	Good control (N = 4763)		Bad control (N = 16,113)	
	Mean	S.D.	Mean	S.D.
Age (year)	63.09	10.59	67.86	10.23
Diagnosis duration (year)	6.77	5.77	10.23	7.57
Haemoglobin, HbA1c (%)	6.44	0.53	8.91	1.42
Random blood glucose, LabRBG (%)	8.46	2.59	12.64	5.74
Triglycerides (mmol/l)	2.15	1.12	2.98	2.31
Cholesterol (mmol/l)	5.02	1.19	5.16	1.20
BMI (kg/m^2)	28.41	8.17	32.45	8.83
Systolic blood pressure (mmHg)	139.93	22.04	147.42	22.74
Diastolic blood pressure (mmHg)	78.55	12.71	81.72	12.56

strategy to evaluate performance and determine the suitable size of features for classification [41,42].

5. Results and discussion

5.1. Feature ranking

FSSMC was used to rank the top 15 features (out of 47) used for the prediction of the patients' diabetes control (Table 3). These features enabled the classifiers to achieve their best performance. Features with higher scores contribute most to the outcome prediction [37].

FSSMC's advantage is its computational efficiency. After selective sampling, 2365 instances were used for feature evaluation, which reduced the number of iterations in further analysis. The reduction rate of sampling instances was 88.7%.

It is well known from clinical studies that type 2 diabetes is a progressive condition with overall blood glucose control deteriorating over time [43,44]. Older patients ($\geq$65) and those who have been diagnosed with type 2 diabetes for longer generally have worse overall control, because blood glucose concentrations tend to increase progressively with age. It is therefore reassuring from the clinical standpoint (and affirms the validity of the data mining techniques used) that 'age' and 'diagnosis-duration' were the features selected as the principle factors in determining whether overall blood glucose control was good or bad.

Improving blood glucose control is a key aim as it reduces the risk of long-term complications. This can be achieved with improved diet (diet treatment), regular physical activity, oral medications (tablet treatment), insulin injections (insulin treatment) or by a combination of these approaches. In type 2 diabetes, as time proceeds, patients generally move in a stepwise fashion through dietary treatment, then oral therapies and eventually need insulin therapy. Despite all of these treatments blood glucose control may continue to deteriorate with time, so it is likely that those on 'insulin-treatment' have the worst overall control, and those on 'diet-treatment' have better control. This again concurs with clinical practice.

Random blood glucose measurements can be taken at any time. Blood glucose level will fluctuate depending on whether the test is carried out before or shortly after a patient has eaten or taken something to drink and is commonly used to evaluate a patient's blood glucose control. It is therefore not surprising that it correlates with HbA1c.

Evidence that genetic factors are important in the aetiology of the condition includes the high rate of concordance for the disease in identical twins; familial aggregation of cases; marked differences in its prevalence in different ethnic groups [4]. Additional evidence shows that genetic factors are important in the pathogenesis of type 2 diabetes [45–47]. Identifying the genetic defects associated with the blood glucose control may provide an opportunity to devise novel methods of treatment and prevention. The presence of 'family histroy' in

Table 3 Control (or otherwise) of the top 15 predictors (out of 47) ranked by FSSMC

Rank	Feature	Explanation	If directly controllable and by clinician or patient
1	Age	Age of patient in years	No
2	Diagnosis duration	Length of time of diagnosis in years	No
3	Insulin treatment	Whether clinician has prescribed insulin (Boolean)	Yes, clinician
4	LabRBG	Laboratory random blood glucose measurement (%)	Laboratory measure
5	Diet treatment	Whether a specific diet has been prescribed expressed in clinical code	Yes, clinician
6	Family history	Whether a patient has a family history of diabetes	No
7	BMI	Body mass index of patient (kg/m^2)	Yes, patient
8	Smoking status	Whether a patient smokes (Boolean)	Yes, patient
9	Glycosuria	The level of sugar in the urine of patient (mmol/l)	Output
10	Complication type	A patient's diabetic related complications expressed in clinical codes, e.g. ICD9	Longer term, clinical aim
11	BPDiastolic	Diastolic blood pressure (mmHg)	Laboratory measure
12	Tablet treatment	Whether clinician has prescribed specific tablet and indication expressed in prescription code	Yes, clinician
13	General proteinuria	The level of the presence of an excess of serum proteins in the urine (mmol/l)	Laboratory measure
14	BPSystolic	Systolic blood pressure (mmHg)	Laboratory measure
15	Lab triglycerides	The level of the presence of triglycerides (mmol/l)	Laboratory measure

the selected factors is therefore consistent with clinical knowledge and experience.

There are strong epidemiological links between type 2 diabetes and obesity [5,48,49], and 'BMI' is a widely used measure of obesity. Longitudinal studies in obesity-prone rhesus monkeys have confirmed that reducing obesity can prevent the development of type 2 diabetes [50]. Obesity is a major predisposing factor to insulin resistance and is also an important obstacle to the effective management of type 2 diabetes. Clinical evidence clearly proves that weight reduction will help improve blood glucose control. In terms of patient controllable factors, this is a highly ranked feature (first) and adds significant weight to the lobby for education to improve lifestyle [51—53].

'Glycosuria' refers to the presence of sugar in the urine. Glucose is present in glomerular filtrate but is reabsorbed by the kidney's proximal tubule. If the blood glucose level exceeds the capacity of the tubules to reabsorb all the glucose present in the glomerular filtrate, the renal threshold is reached and glucose spills into the urine. A finding of glycosuria indicates that the person is hyperglycemic or has a lowered renal threshold for glucose. So clearly, glycosuria will be often present and the degree of it will be elevated if an individual does not control his blood glucose well. It is clinically reasonable therefore that 'glycosuria' should be one of the key predictors for blood glucose classification.

Diabetes is a major cause of heart disease and stroke. According to Vijan and Hayward [54], aggressive blood pressure control may be the most important factor in preventing adverse outcomes in patients with type 2 diabetes. Good care to keep blood glucose under control can prevent the development of hypertension and cardiovascular disease. The control of blood pressure is extremely important in preventing diabetes complications and death [54—56].

The measurement of 'proteinuria' and 'triglycerides' are important laboratory tests. Proteinuria describes a condition in which the urine contains an abnormal amount of protein and this is strongly associated with blood glucose control. Recent clinical trials support the concept that reductions of blood proteinuria correlate with a slowed progression of nephropathy in type 2 diabetes [57]. Triglycerides are the chemical form in which most fat exists in food as well as in the body. They are also present in blood plasma and, in association with cholesterol, make up the plasma lipids. Elevated triglycerides are related to obesity [58] and can be a consequence of untreated or poorly treated diabetes. Better control of body weight will reduce blood glucose level and several studies have suggested the importance of serum triglyceride levels in people with diabetes [59—61].

FSSMC values an attribute by measuring its ability to differentiate the instances from different classes. A relevant attribute has a positive value; vice versa. According to FSSMC, 'drug type' is allocated a negative value, which is expected by clinicians, because of the many different types of drugs used to treat diabetes and diabetic complications only a few would impact on blood glucose control.

The preceding factors can be easily interpreted and support current clinical knowledge, but the diabetic expert was surprised that 'BMI' did not appear in the top five predictors, and this seems to counteract accepted opinion. Similarly, it is also interesting that 'smoking' (which is of course patient controllable) has a high relevance for blood glucose control prediction. Smoking is a major risk factor for macro-vascular complications but is only very loosely associated with adverse metabolic effects leading to poor blood sugar control. However further inspection of the literature shows that there is a body of evidence to suggest that smoking is an independent risk factor [62—64], producing increased macro-vascular morbidity and mortality. Smoking may also be indicative of secondary factors, such as poorer lifestyle choices and the inability to make lifestyle changes.

Thus feature reduction and ranking can provide clinicians with insight into their databases and lead to further understanding of the disease manifestation [65]. FSSMC as a computationally efficient feature selection algorithm can be applied to real world problems and contribute to the progression of data mining in medicine.

5.2. Classification accuracy, sensitivity and specificity

In an attempt to assess the performance of FSSMC as a feature selector on the diabetic data, CA has been used as a preliminary criterion for evaluation. Because it is difficult to estimate the correct number of predictors in feature mining applications [40] (among features with positive values), different sizes of attribute subsets were selected for each of the three classification algorithms (Naïve Bayes, IB1 and C4.5) to find which set gives the best performance. The performance of the models is listed in Table 4. CA measures the proportion of correctly classified test examples, and is used as the principle standard to evaluate the performance of feature selectors and classifiers.

From the table, we can deduce that before feature selection (i.e. 47 variables used), C4.5 had the best performance. However, after feature selection, CA of

Table 4 CA (%) for different sizes of feature subsets (10-CV)

Variable number	Naïve Bayes CA	IB1 CA	C4.5 CA	Average CA
5	84.46	90.96	91.77	89.10
8	86.23	95.26	92.45	91.31
10	88.79	94.21	93.01	92.00
15	87.14	95.04	94.97	92.38
20	85.74	90.55	92.67	89.65
25	83.97	84.31	91.35	86.88
30	83.01	83.45	90.37	85.61
35	82.94	84.20	89.73	85.63
47	78.31	80.13	89.46	82.63
Average	84.52	88.68	91.75	—

Naïve Bayes and IB1 was significantly improved (by more than 10% and 15%, respectively), and IB1 achieved the best accuracy (95.26%). When the predictor number is 10, Naïve Bayes achieved its best result; IB1 obtained the highest CA with 8 predictors; C4.5 performed best while the top 15 features are selected for analysis. On average, when the top 15 variables were selected for classification, the best prediction result can be obtained. The study indicates that feature selection did not affect CA of C4.5 as much as IB1 and Naïve Bayes. Decision tree schemes have an inherent selection mechanism and feature selection provides smaller improvements than other classifiers [40]. IB1 and Naïve Bayes benefited from the reduction of the input parameters. These approaches cannot filter irrelevant or correlated information in the representation and quality of data will affect their performance.

For different classifiers, the best performance was achieved with different feature subsets. The CA improved when correlated and irrelevant features were removed. According to the decision tree constructed by C4.5, the variable 'insulin treatment' was the best predictor for classifying patients' disease control. Among patients, 'age' was the second best predictor for classification. The attributes 'family history' and 'diagnosis duration' are also key for distinguishing the bad blood glucose control patients. The rules generated by C4.5 provide further insight. For example, 'monitor blood' was important according to the decision tree generated by C4.5, which is consistent with the result of FSSMC. This is a reassuring observation, confirming that patients who monitor their blood glucose regularly tend to have better diabetic control. Public health workers may want to explore what educational interventions can be successfully directed to diabetes. It is surprising that in some cases, the patients valued as 'hospital' of attribute 'care type' showed more positive results than those valued as 'general practitioner' and 'shared care'. That is to say, patients treated in hospital clinics may obtain better control than those getting care and advice from both hospital staff and their 'general practitioner' (shared care). That may be due to a more intensive approach to management and lifestyle control in the more specialised hospital-based clinic. However, this is counter intuitive as hospital attending patients could be expected to be more complex and in worse health. This could, of course, be due to secondary illness and not attributed to blood glucose control per se.

In this study, it was more important to detect bad control individuals from the population than to minimize the detection of good control cases. Sensitivity and specificity [66] were used to supplement CA. Table 5 shows the sensitivity and specificity generated by each classifier based on 10-fold cross-validation. IB1 distinguished bad blood glucose control patients from the population (>98%), and C4.5 detected good blood glucose control patients best (>95%). C4.5 had the least difference between sensitivity and specificity and was able to distinguish nearly 95% of bad control cases and 96% of good cases, when the top 15 features were selected. It can be seen from the results that the

Table 5 Sensitivity and specificity for different feature subsets (sensitivity/specificity) based on 10-CV

Feature number	Naïve Bayes	IB1	C4.5	Average
5	88.12/72.12	97.82/67.78	93.69/85.28	93.21/75.06
8	90.07/73.24	98.80/83.28	93.73/88.12	94.20/81.55
10	91.43/79.86	97.94/81.59	94.05/89.48	94.47/83.64
15	88.28/83.24	98.33/83.92	94.74/95.74	92.38/92.74
20	90.98/74.15	96.35/70.93	93.40/90.20	93.58/78.43
25	89.66/72.48	89.85/65.57	91.76/89.96	90.42/76.00
30	89.71/60.35	88.24/67.25	91.02/88.17	89.66/71.92
35	89.31/61.40	89.47/66.38	89.96/88.95	89.58/72.24
47	85.78/53.05	87.77/53.92	89.64/88.85	87.13/65.27
Average	89.26/69.99	93.84/71.18	92.44/89.42	—/—

Table 6 Feature selection performance comparison using UCI databases and the UCHT diabetes dataset

Data set	Instance number (s)	InfoGain time (s)	ReliefF time (s)	FSSMC instance number after reduction (%)	FSSMC time (s)
Breast	699	0.16	2.05	45 (94)	0.15
Credit	690	1.15	3.58	159 (77)	1.20
Pima Indian diabetes	768	1.26	2.63	240 (69)	1.34
Glass	214	0.44	0.56	80 (63)	0.67
Heart	294	0.22	0.32	39 (87)	0.41
Hypothyroid	3,772	1.47	102	415 (89)	12.8
Kr-vs-Kp	3,196	0.73	156	496 (85)	18.0
Letter	20,000	2.62	2914	1858 (91)	377
Mushroom	8,124	0.65	446	89 (99)	5.86
Soybean	683	0.34	5.92	109 (84)	1.73
Splice	3,190	0.54	164	254 (92)	25.7
Waveform	5,000	1.96	445	457 (91)	67.4
UCHT diabetes	20,876	4.16	3989	2365 (89)	793

feature selection technique is also able to influence the value of sensitively and specificity, i.e. using more relevant information can result in improvement of sensitivity and specificity. Naïve Bayes and IB1 had good sensitivity with respect to the particular number of features but with relative low CA.

5.3. Computational efficiency

FSSMC as a feature selector improves computational efficiency whilst maintaining CA when compared to ReliefF. Table 6 shows a comparison using well-known data sets drawn from the University of California Irvine (UCI) repository of machine learning databases [67]. Since the relevant features are unknown in advance for these data sets, the performance of classification algorithms indicates the effect of Information Gain, Relief, and FSSMC in selecting useful features on various types of data sets.

Information Gain is the fastest method as it does not calculate distance between two instances and it pays more attention to features instead of instances. ReliefF and FSSMC ranks the ability of a feature to discriminate between classes by calculating the similarity of instances, which is more time consuming. FSSMC decreases the processing time of ReliefF in large datasets (instances >3000) such as 'hypothyroid', 'Kr-vs-Kp', 'letter', 'mushroom', 'splice', 'waveform' and UCHT diabetes. But for small data sets, 'glass' and 'heart' (instances <300), FSSMC requires the longest processing time. For FSSMC, on average, only 14.6% of instances are selected as the start data points for feature analysis, which leads to the significant improvement in efficiency, compared to ReliefF. For these data sets CA is maintained, Table 7. This shows CA derived for Naïve Bayes using 10-fold cross-validation. Similar results derived for C4.5 and IB1 classifiers (not shown) also indicate that FSSMC maintains CA.

Table 7 CA comparison for Information Gain, ReliefF and FSSMC using UCI databases and UCHT diabetes dataset (CA averaged over 10-fold)

Data set	CA for Naïve Bayes and Information Gain	CA for Naïve Bayes and RefiefF	CA for Naïve Bayes and FSSMC
Breast	96.3	96.7	96.7
Credit	79.0	84.2	84.1
Pima Indian diabetes	76.2	76.0	76.9
Glass	48.6	48.1	48.1
Heart	83.3	84.0	84.9
Hypothyroid	95.3	95.3	95.3
Kr-vs-Kp	88.0	88.9	88.9
Letter	64.2	64.6	64.6
Mushroom	97.8	98.5	98.7
Soybean	92.8	92.8	92.8
Splice	95.6	95.6	95.6
Waveform	80.2	80.2	80.1
Ulster diabetes	78.3	80.2	81.0

Table 8 Decision tree generation and processing time for C4.5 (10-CV) and processing time for IB1 (10-CV), as a function of increasing feature number

Feature number	C4.5			IB1
	Leaf number	Tree size	Processing time (min, s)	Processing time (min, s)
5	43	52	43 s	5 m 10 s
8	51	68	50 s	8 m 45 s
10	70	75	56 s	20 m 50 s
15	74	87	1 m 15 s	33 m 24 s
20	93	95	1 m 37 s	45 m 32 s
25	84	109	1 m 54 s	56 m 7 s
30	87	116	2 m 16 s	63 m 45 s
35	93	130	1 m 57 s	80 m 5 s
47	80	162	3 m 58 s	118 m 8 s

Feature selection reduces the dimension of datasets and enhances the computational efficiency of classification algorithms. Table 8 shows the influence of dimension reduction on the processing time of C4.5 and IB1. The processing time reduces with the decreasing number of attributes. For IB1, processing time shortened to 7.6% original processing time (without feature selection) after removing irrelevant features.

The size of the generated decision tree also reduces with the reduction of attributes. However, C4.5 generates 93 leaves with 20 attributes, more than 80 leaves with 47 attributes. The reason is that decision trees are required to fit the classes in the data; when some attributes are removed from the data set, it is possible that more 'branches' will be generated during the learning procedure. Smaller trees are easier to interpret, relieve the burden of pruning, and reduce the probability of over-fitting.

6. Conclusion

FSSMC was applied to identify the key factors affecting blood glucose control among type 2 diabetic patients. On average, classifiers were able to achieve their best performance when the top fifteen features were selected. FSSMC enhanced the sensitivity and specificity of classifiers in these cases. In this research, the major discriminative factors were found to be: 'age', 'diagnosis duration', 'insulin treatment', 'random blood glucose' and 'diet treatment'. While age and duration of diagnosis are variables which cannot be controlled, blood glucose levels can be positively influenced by management of the disease. The clinician can suggest dietary adjustments, new drug therapies and lifestyle changes. The diabetic specialist was encouraged by the fact that the data-driven analysis in this study was able to verify by importance of rank much of the knowledge which has been researched for over 50 years in the medical domain using epidemiological studies, laboratory investigation and inductive reasoning by clinicians. Although this is not surprising, as the domain knowledge obviously influences the data collected, the ranking of predictors (Table 3) provides a verification of the feature selection approach adopted, and suggests the efficacy of feature selection and data mining in large clinical databases.

Naïve Bayes, IB1 and C4.5 were used to predict diabetes control. The performance of IB1 and Naïve Bayes were significantly improved by applying FSSMC feature selection. This may be due to the elimination of noisy and irrelevant features that may mislead the learning process. The results show that the processing times of IB1 and C4.5 classifiers were reduced after feature selection, and the tree generated by C4.5 was reduced in size. The models provided a best predictive accuracy of 95% and sensitivity of 98%. Additionally, some novel knowledge was mined from the data, such as the importance of 'smoking' (ranked eighth overall) on outcome, the relationship between 'care type', and the need for intensive home blood monitoring to control diabetes. These findings contribute to knowledge in the application domain. In spite of the requirement for independent verification of the knowledge obtained, this work supports the further use of data mining approaches in understanding diabetes and in the general medical domain, which is suffering from a data explosion due to enhanced procedures for recording data.

The experimental results verify FSSMC as a practical feature estimator for classifiers in the diabetic domain. FSSMC preserved the accuracy of the full set classifier, while using the selected available variables and significantly improved the computational efficiency of two important classification algorithms.

Although IB1 appears to be the best for classification, and Naïve Bayes processes the data fastest,

C4.5 is the most stable classifier, with the highest precision and the best balance between sensitivity and specificity. There are three factors which suggest that these techniques will become more important over time:

1. The number of follow-up interventions on existing patients will provide a richer dataset.
2. The number of records will increase as more patients are examined and their data are stored.
3. The quality of the data will be improved as health care professionals become more conscientious in recording data and the importance of the dataset is realized.

Overall there was high concordance between the features selected using FSSMC and the factors anticipated as being important by the diabetes expert. The models' predictive performance and the clinical relevance of the features selected suggest that decision support and prediction in routine practice could be achievable. The most important factors identified such as 'age' and 'diagnosis duration' are beyond the control of the treating physician and stress the need for prevention and public health education. This research supports the evidence based medicine paradigm which is widely accepted as the basis for better medical practice.

References

[1] Gan D, editor. Diabetes atlas, 2nd ed. Brussels: International Diabetes Federation; 2003. http://www.eatlas.idf.org/webdata/docs/Atlas%202003-Summary.pdf (accessed June 19, 2007).

[2] Alberti K, Zimmet P. Definition, diagnosis and classification of diabetes mellitus and its complications. Part 1. Diagnosis and classification of diabetes mellitus—provisional report of a WHO Consultation. Diabetic Med 1998;15:539—53.

[3] Guthrie RA, Guthrie DW, editors. Nursing management of diabetes mellitus. 5th ed., New York: Springer Publishing; 2002.

[4] Pinhas-Hamiel O, Zeitler P. Acute and chronic complications of type 2 diabetes mellitus in children and adolescents. Lancet 2007;369:1823—31.

[5] Pickup JC, Williams G, editors. Textbook of diabetes. 3rd ed., Oxford: Blackwell Science; 2003.

[6] Lorig K, Holman H. Self management education: history, definition and outcomes and mechanisms. Ann Behav Med 2003;26(1):1—7. doi:10.1207/S15324796ABM2601_01.

[7] Smith R. Improving the management of chronic disease. Br Med J 2003;327. doi:10.1136/bmj.327.7405.12.

[8] Department of Health. Supporting people with long term conditions: an NHS and social care model to support local innovation and integration. London: Department of Health; Crown copyright 2005.

[9] Department of Health. Self care: a real choice. London: Department of Health; Crown copyright 2005.

[10] Nissen SE, Wolski K. Effect of rosiglitazone on the risk of myocardial infarction and death. N Engl J Med 2007;365. doi:10.1056/NEJMoa072761.

[11] Dash M, Liu H. Consistency-based search in feature selection. Artif Intell 2003;151:155—76.

[12] Lavrac N. Data mining in medicine: selected techniques and applications. In: Proceedings of the second international conference on the practical application of knowledge discovery and data mining. London: The Practical Applications Company; 1998. p. 11—31.

[13] Mitchell M, editor. Machine learning. New York: McGraw-Hill; 1997.

[14] Martin B. Instance-based learning: nearest neighbour with generalisation. PhD thesis. Hamilton, New Zealand: Department of Computer Science, University of Waikato; 1995.

[15] Lewis D, Gale W. A sequential algorithm for training text classifiers. In: Croft BW, Rijsbergen CJ, editors. Proceedings of the seventeenth annual ACM-SIGIR conference on research and development in information retrieval. Springer-Verlag; 1994. p. 3—12.

[16] Rish I, Hellerstein J, Thathachar J. An analysis of data characteristics that affect Naïve Bayes performance. New York. IBM Technical Report; 2002. http://www.research.ibm.com/PM/icml01.pdf (accessed June 19, 2007).

[17] Topon KP. Gene expression based cancer classification using evolutionary and non-evolutionary methods. Technical Report No. 041105A1. Japan: Department of Frontier Informatics, The University of Tokyo; 2004.

[18] Cornforth D, Jelinek H, Peichl L. Fractop: a tool for automated biological image classification. In: Sarker, McKay, Gen, Namatame, editors. Proceedings of the sixth Australia—Japan joint workshop on intelligent and evolutionary systems. 2002. p. 141—8.

[19] Aires R, Manfrin A, Aluisio S, Santos D. Which classification algorithm works best with stylistic features of Portuguese in order to classify web texts according to users needs? Technical Report NILC-TR-04-09. Brasil: University de Sao Paulo; 2004.

[20] Hall M. Correlation-based feature selection for machine learning. PhD thesis. Hamilton, New Zealand: Department of Computer Science, University of Waikato; 1999. http://www.cs.waikato.ac.nz/~mhall/thesis.pdf (accessed June 19, 2007).

[21] Inza I, Sierra B, Blanco R, Larranaga P. Gene selection by sequential search wrapper approaches in microarry cancer class prediction. J Intell Fuzzy Syst 2002;12(1):25—32.

[22] Hall M, Holmes G. Benchmarking attribute selection techniques for discrete class data mining. IEEE Trans Knowledge Data Eng 2003;15:1437—47.

[23] Sierra B, Lazkano E. Probabilistic-weighted *k*-nearest neighbour algorithm: a new approach for gene expression-based classification. Knowledge-Based Intell Inf Eng 2003;932—9.

[24] Su CT, Yang CH, Hsu KH, Chiu WK. Data mining for the diagnosis for type II diabetes from three-dimensional body surface anthropometrical scanning data. Comput Math Appl 2006;51:1075—92.

[25] Huang Y, McCullagh PJ, Black ND. Feature selection via supervised model construction. In: Bramer M, editor. Proceedings of the 4th IEEE international conference on data mining. 2004. p. 411—4.

[26] Kononenko I. Estimating attributes: analysis and extension of relief. In: Proceedings of the seventh European conference in machine learning. Springer-Verlag; 1994. p. 171—82.

[27] Demsar J, Zupan B, Aoki N, Wall M, Granchi T, Beck J. Feature mining and predictive model construction from severe trauma patient's data. Int J Med Inf 2001;63:41—50.

[28] Kononenko I, Simec E. Induction of decision trees with RELIEFF. In: Proceedings of ISSEK workshop on mathematical and statistical methods in artificial intelligence. New York: Springer; 1995. p. 199—220.
[29] Robnik M, Kononenko I. Theoretical and empirical analysis of ReliefF and RReliefF. Mach Learn 2003;53:23—69.
[30] Fayyad U, Piatesky-Shapiro G, Smyth P, editors. Advances in knowledge discovery and data mining. AAAI/MIT Press; 1996.
[31] Kauderer K, Mucha H, editors. Classification, data analysis and data highways. New York: Springer-Verlag; 1997.
[32] Schohn G, Cohn D. Less is more: active learning with support vector machines. In: Pat Langley, editor. Proceedings of the seventeenth international conference on machine learning. Morgan Kaufmann; 2000. p. 839—46.
[33] Roy N, McCallum A. Toward optimal active learning through sampling estimation of error reduction. In: Brodley CE, Pohoreckyj Danyluk A, editors. Proceedings of the eighteenth international conference on machine learning. Morgan Kaufmann; 2001. p. 441—8.
[34] Liu H, Motoda H, Yu L. A selective sampling approach to active feature selection. Artif Intell 2004;159:49—74.
[35] Aha D, Kibler D, Albert M. Instance-based learning algorithms. Mach Learn 1991;6:37—66.
[36] Kantardzic M, editor. Data mining: concepts, models, methods, and algorithms. New Jersey: Wiley-IEEE Press; 2002.
[37] Demsar J, Zupan B, Aoki N, Wall MJ, Granchi TH, Beck JR. Feature mining and predictive model construction from severe trauma patient's data. Int J Med Inf Elsevier Science 2001;63:41—50.
[38] Molina L, Belanche L, Nebot A. Feature selection algorithms: a survey and experimental evaluation. In: Proceeding of IEEE international conference on data mining, IEEE. 2002. p. 306—13.
[39] van Bemmel J, Musen M, editors. Handbook of medical informatics. New York: Springer; 1997.
[40] Perner P. Improving the accuracy of decision tree induction by feature pre-selection. Appl Artif Intell 2001;15(8):747—60.
[41] Grzymala-Busse J. Data mining in bioinformatics. Technical Report. USA; University of Kansas; 2003.
[42] Hall L, Collins R, Bowyer K, Banfield R. Error-based pruning of decision trees grown on very large data sets can work. In: Proceedings of 14th IEEE international conference on tools for artificial intelligence; 2002. p. 233—8.
[43] Bennett P. Epidemiology of diabetes mellitus. In: Rifkin H, Porte D, editors. Ellenberg and Rifkin's diabetes mellitus. New York: Elsevier; 1990. p. 363—77.
[44] Croxson S, Burden A, Bodlington M, Bostha J. The prevalence of diabetes in elderly people. Diabetic Med 1991;8:28—31.
[45] Newman B, Selby J, King M. Concordance for type 2 diabetes mellitus (NIDDM) in male twins. Diabetiologia 1987;30: 763—8.
[46] Knowler W, Pettitt D, Saad M. Diabetes mellitus in the pima Indians: Incidence, risk factors and pathogenesis. Diabetes Metab Rev 1990;6:1—27.
[47] Harris M. Epidemiological correlates of NIDDM in Hispanics, Whites, and Blacks in the US population. Diabetes Care 1991;14:639—48.
[48] Marcovecchio M, Mohn A, Chiarelli F. Type 2 diabetes mellitus in children and adolescents. J Endocrinol Investig 2005;28:853—63.
[49] Wong T, Barr E, Tapp R, Harper C, Taylor H, Zimmet P, et al. Retinopathy in persons with impaired glucose metabolism: the Australian diabetes obesity and lifestyle (AusDiab) study. Am J Ophthalmol 2005;140:1157—9.
[50] Hansen B, Bodkin N. Primary prevention of diabetes mellitus by prevention of obesity in monkeys. Diabetes 1993;42: 1809—14.
[51] Brug J, Campbell M, van Assema P. The application and impact of computer generated personalized nutrition education: a review of the literature. Patient Educ Counsel 1999;36:145—56.
[52] Diabetes Prevention Program Research Group. Reduction in the incidence of Type II diabetes with lifestyle intervention or metformin. N Engl J Med 2002;346(6):393—403.
[53] Franz MJ. The answer to weight loss is easy—doing it is hard! Clin Diabetes 2001;19(3):105—9.
[54] Vijan S, Hayward RA. Treatment of hypertension in Type 2 diabetes mellitus: blood pressure goals, choice of agents, and setting priorities in diabetes care. Ann Intern Med 2003;138:593—602.
[55] The American College of Physicians. Blood pressure control in people with Type 2 diabetes mellitus: recommendations from the American College of Physicians. Ann Intern Med 2006;138:1—70.
[56] Snow V, Weiss KB, Mottur-Pilson C. The evidence base for tight blood pressure control in the management of Type 2 diabetes mellitus. Ann Intern Med 2003;138:587—92.
[57] Bakris G, Weir M, DeQuattro M, McManhon F. Effects of an ace inhibitor/calcium antagonist combination on proteinuria in diabetic nephropathy. Kidney Int 1998;54:1283—9.
[58] Cheraskin E. The breakfast/lunch/dinner ritual. J Orthomol Med 1993;8:6—10.
[59] West K, Ahuja M, Bennett B, Czyzyk A, DeAcosta O, Fuller J. The role of circulating glucose and triglyceride concentrations and their interactions with other 'risk factors' as determinants of arterial disease in nine diabetic population samples from the who multinational study. Diabetes Care 1983;6:361—9.
[60] Standl E, Stiegler H, Janka H, Mehnert H. Risk profile of macrovascular disease in diabetes mellitus. Diabetes Metab 1988;14:505—11.
[61] Fontbonne A, Thibult N, Eschwege E, Ducimetiere P. Body fat distribution and coronary heart disease mortality in subjects with impaired glucose tolerance or diabetes mellitus: the paris prospective study 15-year follow-up. Diabetologia 1992;35:464—8.
[62] Rimm E, Chan J, Stampfer M, Colditz G, Willett W. Prospective study of cigarette smoking, alcohol use, and the risk of diabetes in men. Br Med J 1995;310:555—9.
[63] Wannamethee, Shaper SA, Perry I. Smoking as a modifiable risk factor for type 2 diabetes in middle-aged men. Diabetes Care 2001;24:1590—5.
[64] Sairenchi T, Iso H, Nishimura A, Hosoda T, Irie F. Cigarette smoking and risk of type 2 diabetes mellitus among middle-aged and elderly Japanese men and women. Am J Epidemiol 2004;160:158—62.
[65] Chen M, Han J, Yu P. Data mining: an overview from a database perspective. IEEE Trans Knowledge Data Eng 1996;8:866—83.
[66] Veropoulos K, Campbell C, Cristianini N. Controlling the sensitivity of support vector machines. In: Proceedings of the international joint conference on artificial intelligence (IJCAI) Workshop support vector machines; 1999. p. 55—60.
[67] Newman DJ, Hettich S, Blake CL, Merz CJ. UCI repository of machine learning databases. Irvine, CA: University of California, Department of Information and Computer Science; 1998. http://www.ics.uci.edu/∼mlearn/MLRepository.html (accessed June 19, 2007).
[68] Crone S, Lessmann S, Stahlbock R. Empirical comparison and evaluation of classifier performance for data mining in customer relationship management. In: Wunsch D, et al., editors. Proceedings of the international joint conference on neural networks, IJCNN'04. 2004. p. 443—8.