British Journal of Applied Science & Technology
4(33): 4591-4606, 2014
ISSN: 2231-0843

SCIENCEDOMAIN *international*
www.sciencedomain.org

Novel Data Mining Techniques for Incomplete Clinical Data in Diabetes Management

Herbert F. Jelinek[1], Andrew Yatsko[2], Andrew Stranieri[2] and Sitalakshmi Venkatraman[3*]

[1]Centre for Research in Complex Systems and School of Community Health, Charles Sturt University, P.O. Box 789, Albury, NSW 264, Australia.
[2]Centre for Informatics and Applied Optimisation, Federation University, P.O. Box 663, University Drive, Mt Helen Vic 3350, Australia.
[3]Department of Higher Education - Business (IT), Northern Melbourne Institute of TAFE, 77-91 St Georges Rd, Preston Victoria 3072, Australia.

Authors' contributions

This work was carried out in collaboration among all authors. Author HFJ designed the study and provided the clinical data. Author AY performed the data mining of the datasets with the guidance of author AS and wrote the first draft of the manuscript. Author SV managed the relevant contemporary literature studies and developed the final manuscript with refined analyses of the study. All authors read and approved the final manuscript.

Article Information

DOI: 10.9734/BJAST/2014/11744
Editor(s):
(1) Selvakumar Subbian, Laboratory of Mycobacterial Pathogenesis and Immunity, Public Health Research Institute (PHRI) at Rutgers Biomedical and Health Sciences, Newark, USA.
Reviewers:
(1) Anonymous, King Abdulaziz University, Saudi Arabia.
(2) Maliheh Dadgarmoghaddam, Mashhad University of Medical Sciences, Iran.
(3) D. Senthilkumar, Computer Science and Engineering, Bharathidasan Institute of Technology, Anna University, India.
Complete Peer review History: http://www.sciencedomain.org/review-history.php?iid=670&id=5&aid=6128

Original Research Article

Received 1st June 2014
Accepted 19th August 2014
Published 16th September 2014

ABSTRACT

An important part of health care involves upkeep and interpretation of medical databases containing patient records for clinical decision making, diagnosis and follow-up treatment. Missing clinical entries make it difficult to apply data mining algorithms for clinical decision

**Corresponding author: E-mail: SitaVenkat@nmit.edu.au;*

support. This study demonstrates that higher predictive accuracy is possible using conventional data mining algorithms if missing values are dealt with appropriately. We propose a novel algorithm using a convolution of sub-problems to stage a super problem, where classes are defined by *Cartesian Product* of class values of the underlying problems, and *Incomplete Information Dismissal* and *Data Completion* techniques are applied for reducing features and imputing missing values. Predictive accuracies using Decision Branch, Nearest Neighborhood and Naïve Bayesian classifiers were compared to predict diabetes, cardiovascular disease and hypertension. Data is derived from Diabetes Screening Complications Research Initiative (DiScRi) conducted at a regional Australian university involving more than 2400 patient records with more than one hundred clinical risk factors (attributes). The results show substantial improvements in the accuracy achieved with each classifier for an effective diagnosis of diabetes, cardiovascular disease and hypertension as compared to those achieved without substituting missing values. The gain in improvement is 7% for diabetes, 21% for cardiovascular disease and 24% for hypertension, and our integrated novel approach has resulted in more than 90% accuracy for the diagnosis of any of the three conditions. This work advances data mining research towards achieving an integrated and holistic management of diabetes.

Keywords: Data mining; missing value imputation; diabetes management; classifiers; diagnosis accuracy.

1. INTRODUCTION

There has been a growing interest in understanding the applications of data mining for clinical decision support in metabolic syndrome and diabetes mellitus type 2 due to its clinical complexity and its association with an increasing risk for heart disease and stroke, and hypertension [1-3]. Many studies have consistently reported suboptimal diabetes control outcomes despite improvements in clinical information and self-management support systems that have been recently established [4-9]. This is in part due to missing clinical data which reduces the information content for accurate decision making. The rationale of the study is that, given appropriate algorithms for imputing missing values, the accuracy of diagnosis based on clinical data should improve. The prime objective of this study is to propose novel imputation techniques in data mining to improve the predictive accuracy of diabetes, hypertension and cardiovascular disease (CVD) from clinical data.

A naïve approach to dealing with a record having one attribute value missing involves deleting the entire record. However, the removal of records may leave too few examples covering specialized subsets of data and compromises the analytics. Another approach involves restructuring the dataset into smaller sets, where each new set has minimal missing values [10]. However, the decomposition of the data mining task into a series of smaller interconnected tasks can impose a structure that reduces the associations that can be discovered across all attributes. An alternative is to substitute the missing values with imputed values. Experts familiar with diabetes can conceivably impute missing values manually but this is rarely practical for assembling a large diabetes dataset as the amount of missing data may be extensive and the method is subjective and therefore not repeatable.

Common approaches to imputation involve filling in missing values with the mean values of corresponding attributes. This is computationally easy to apply but can compromise generalisation because the record with the missing values may not be representative of all records. However, if the mean is calculated solely from those records that have the same class value as the record with the missing value, predictive accuracy can be expected to increase.

The approach presented in this study involves two strategies we call *Incomplete Information Dismissal* and *Contextual Data Completion by Mode*. *Incomplete Information Dismissal* involves eliminating features and records that do not contribute to a classification. This involves a formally specified feature ranking algorithm. In [11], each class is considered as a cluster of data points in the problem space. The cluster centre represents a signature for the class. A new instance is classified by selecting the cluster centre closest to the point representing the new instance. The least important features for a class are those that cause data points to be closer to centres of other classes. With nominal/discrete data, it is more convenient to use feature ranking based on association measures, particularly Information Gain (IG). However, the approach described in [11] is conceptually the same.

The *Data Completion* involves the imputation of missing values. In this approach, a set of points closest to the missing value point are selected, so that the class was the same. Each possible replacement value is evaluated to identify the replacement that will make the missing value instance closest to the others. This approach is dependent on other missing values, so the algorithm iterates through successive imputations. Other approaches to missing value imputation involve use of a classification algorithm where the data set is prepared so that the class is the attribute with the missing value. A classification algorithm such as a decision tree induction exemplified by C4.5 is trained using records with known values and run to predict the attribute's missing value. Imputation using classification algorithms in this way is computationally expensive because a model for each attribute must be trained. Values imputed by our algorithm involving the *Data Completion* approach are finely tuned to more likely lead to high predictive accuracies when the dataset completed with imputed values is used for classification. In addition, we propose a novel integrated approach where the classes are defined by *Cartesian Product* of class values for underlying problems, namely diabetes, cardiovascular disease (CVD) and hypertension so as to set up a super problem that would classify diabetes, CVD and hypertension at the same time - a step toward a holistic management of diabetes.

Some diabetic prediction studies conducted recently [12] have achieved good classification accuracies with a weighted sum approach compared to other data mining models such as logistic regression and neural networks. However, the highest accuracy level achieved to predict diabetes or prediabetes was only 77.87% based solely on traditional risk factors and results are not presented in a clinical useful way. This paper reports that common data mining algorithms can demonstrate higher accuracies for classification of diabetes, CVD and hypertension if incomplete clinical data is managed appropriately by including a two-stage approach of *Incomplete Information Dismissal* and *Data Completion* with convolution (*Cartesian Product*) of class values of the three problems.

1.1 Background

Diabetes Mellitus (DM) is a metabolic disorder of multiple etiology. It is characterized by chronic hyperglycemia and disturbances of carbohydrate, fat, and protein metabolism resulting from defects of insulin secretion, insulin action, or a combination of both [7-8,13]. The most common types of diabetes are type 1 diabetes (insulin sensitive due to cellular-mediated autoimmune destruction of the β-cells of the pancreas in 90% of cases and is idiopathic in 10% of cases), and type 2 diabetes (insulin insensitive due to pancreatic β-cell dysfunction and insulin resistance) [14]. Uncontrolled diabetes leads to a risk of microvascular (retino-, nephro- and neuropathy) and macrovascular (cardio-, cerebrovascular and peripheral vascular) health problems. Diabetes is increasingly reported as a primary cause of death, and the risk of death for people with diabetes is twice as high as for the general population with similar age [15].

While many clinical tests consider various standard factors such as age, blood pressure (BP), age at clinical onset, body weight, family history, urinary blood sugar levels and ketones, there is much more investigation required into how traditional and emerging biomarkers combine for optimal control of diabetes [16,17]. As clinical tests become more sophisticated, there is a huge amount of data collected from each patient on various attributing factors including inflammatory, oxidative stress and genetic biomarkers that need to be integrated into a comprehensive clinical decision making model. However, analysis may be hampered or may not give the correct results when information is missing.

The way a data mining algorithm presents results for interpretation by an analyst depends, to a large extent, on the algorithm. The ID3 classifier described by [18] generates a decision tree where nodes are attributes, arcs represent possible values, and leaf nodes represent data to be classified, as illustrated in Fig. 1. Decision trees can readily be understood by clinicians, however decision trees generated from many variables are typically very large and cannot easily be interpreted [19]. Fig. 2 depicts the weighted sum visualization (AWSum) output [20]. The horizontal bars depict the influence a pair of features has on a classification of diabetes on the right side of Fig. 2 compared with one of no-diabetes on the left side. A clinician can readily ascertain that the propensity toward diabetes for females with a cardiovascular risk of 15-20% is quite high despite gender and CVD risk category not contributing on their own to this conclusion.

Although data mining algorithms that present data visually in ways that enhance insight, as exemplified by Figs. 1 and 2, are emerging, their usefulness is limited by the extent to which missing values impede the classification.

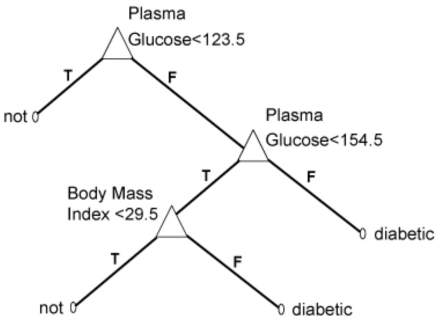

Fig. 1. Conventional decision tree

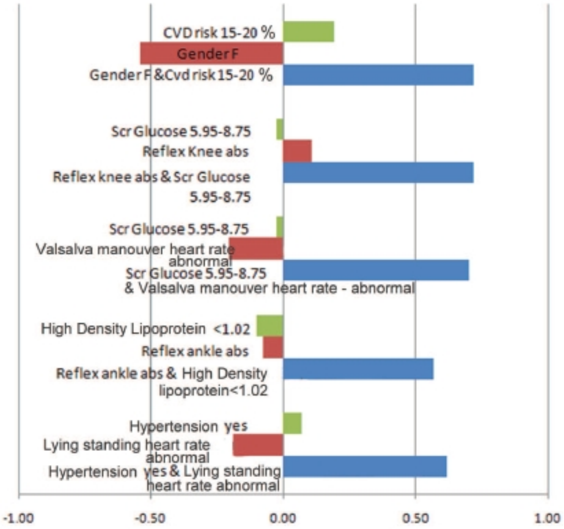

Fig. 2. Weighted sum visualisation

1.2 Missing Value Imputation Techniques

Most techniques for the imputation of missing values treat numerical and categorical attributes differently, so the same technique cannot be used when both numerical and categorical features are present. A review of the imputation of numerical values including techniques based on the Nearest Neighbor approach, Linear Interpolation, Cubic Spline Interpolation and Neural Networks is provided by [21,22]. In general, these approaches use the numerical values in a distance function that represents the degree of similarity between records. In [23,24], approaches based on modeling of a data probability density function using kernels such as Gaussian were evaluated, allowing the relationship between attributes to be exploited. In this connection, [25] discusses the Expectation Maximization algorithm (EM) introduced by the authors of [26], which relies on the data mean and covariance matrix. Generally, a model of data can be fit to a sample. Many statistical software packages (SPSS, SAS, to name a few) implement derivatives of EM and the alternative technique of Multiple Imputation (MI) from the authors of [26]. A review of MV imputation methods in classification is also given in [27]. A more recent missing value imputation technique using an entropy based decision tree algorithm provides better results for datasets having higher correlations among attributes [28]. However, attributes in medical datasets may not necessarily be highly correlated and these methods are therefore still not optimal.

In dealing with incomplete medical data, some researchers have attempted to use missing value imputation with general practice data [29-31], and longitudinal dietary data [32,33]. However, pitfalls in adopting a single technique have been reported by [34]. In [35], less than forty studies that had applied data mining techniques to diabetes were found in a recent systematic review of the literature. The relative paucity of studies can be explained by noting that the performance of an effective data mining study in diabetes requires the assembly of an appropriate dataset, comprehensive approaches for preparing the data for mining from both discrete and continuous data, and presentation of results in a manner that can be meaningfully understood by diabetes analysts.

As clinical tests become more sophisticated and numerous, there is a huge amount of data collected from each patient on various attributing factors. Virtually all patient datasets are replete with missing values for various reasons: equipment used for measurements sometimes malfunctions; a value could be missing because data was never collected – eg HbA1c test was never obtained or the patient was unable to complete the test. In other situations, the value was simply not available at the time of collection but may have become available subsequent to the assembly of the dataset.

This paper addresses the issues of missing values (MV) in clinical data sets, and as a first step proposes novel data mining techniques to deal with, and impute missing values to provide an improved and reliable diagnosis and management of diabetes.

2. MATERIALS AND METHODS

In this section, we describe how clinical datasets were collected and organized, and the proposed methods adopted for the experimental study. We provide below details of how our approach discretizes continuous variables to nominal values and deals with missing values MVs using 3 classifiers: Decision Tree (DT), Nearest Neighbor (NN) and Naïve Bayesian (NB).

2.1 Clinical Datasets

The dataset used in this study is derived from the Diabetes Screening Complications Research Initiative (DiScRi) conducted at a regional Australian university [36]. It is a diabetes complications screening program in Australia where members of the general public participate in a comprehensive health review. The DiScRi community screening concentrates on diabetes, cardiovascular disease and hypertension as a triad of diseases. There are no explicit groupings for retinopathy and neuropathy provided in the database and therefore these were not considered in the current investigation. The screening clinic has been collecting data over ten years and includes over one hundred features such as demographics, socio-economic variables, education background and clinical variables. Clinical variables included blood glucose level, HbA1c, cholesterol profile, inflammatory and oxidative stress markers, other medical history, body mass index, peripheral vascular function, and ECG derived variables. Data on 273 attributes from approximately 2500 attendances of nearly 900 patients have been collected in recent years. The dataset has been used in several data mining applications [37-39]. Application of the Data-driven Decision Guidance Management System (DD-DGMS) approach to this dataset is discussed in the following subsections. Currently, the project is still continuing to collect data, and therefore the dataset is not yet made public until completion of the project.

The database of 2429 records underwent compression to instantiate patients instead of attendances. The latest data in chronological order was used to initialize patient records. Any MVs in 102 applicable attributes were sourced from previous attendances leading to approximately one-fourth of MVs being restored. Altogether, the 824 instances included patients not diagnosed with diabetes mellitus of any type (594), those diagnosed with T2DM (211) and T1DM (19). Type 1 instances were later excluded as the feature-set was not discriminating T1DM enough from control and T2DM, and due to the small T1DM sample. Fig. 3 provides a typical dataset snapshot of the attributing factors collected from the clinical trials of patients diagnosed with diabetes, cardiovascular disease and hypertension. The database includes patients' demographic information that form pre-determined data along with the main clinical data as well as derived data.

The diabetes dataset assembled for this study contained 97 attributes, of which 65% were incomplete, and 805 instances, of which none was complete, with 32% values missing across all data, as appears in Table 1. The class structure for the current classification problem is shown in Table 2. For example in row one, Type II Diabetes is the diagnostic class (Class 1), anything else is in Class 0.

In this work, with many nominal attributes in the data, we adopt an approach based on continuous attributes discretized.

2.1.1 Even frequency discretization

We discretize any continuous attribute using an algorithm we call *Even Frequency Discretization*, so methods that work with all-nominal data could be applied. The technique is an interpretation of the Fixed Frequency method [40]. Real-valued data is distributed into intervals to accumulate frequencies of corresponding discrete values. The frequencies are targeted to be as even as possible. Due to the limited precision in obtaining data, a value, even though perceived as unique, may repeat. The number of intervals is same for all attributes discretized. The adopted approach has proved being reliable in many applications.

Table 1. Diabetes dataset

Attributes			Instances		Values
All	**Numerical**	**Incomplete (%)**		**Incomplete (%)**	**Missing (%)**
97	56	65	805	100	32

Table 2. Data subdivision by class

Problem	Classes (%)		
	Unknown	**0**	**1**
Diabetes mellitus	0	74	26
Cardiovascular disease	28	54	18
Hypertension	20	32	48

'Unknown' – instances not having class attribute set; '0' – control class; '1' – diagnostic class

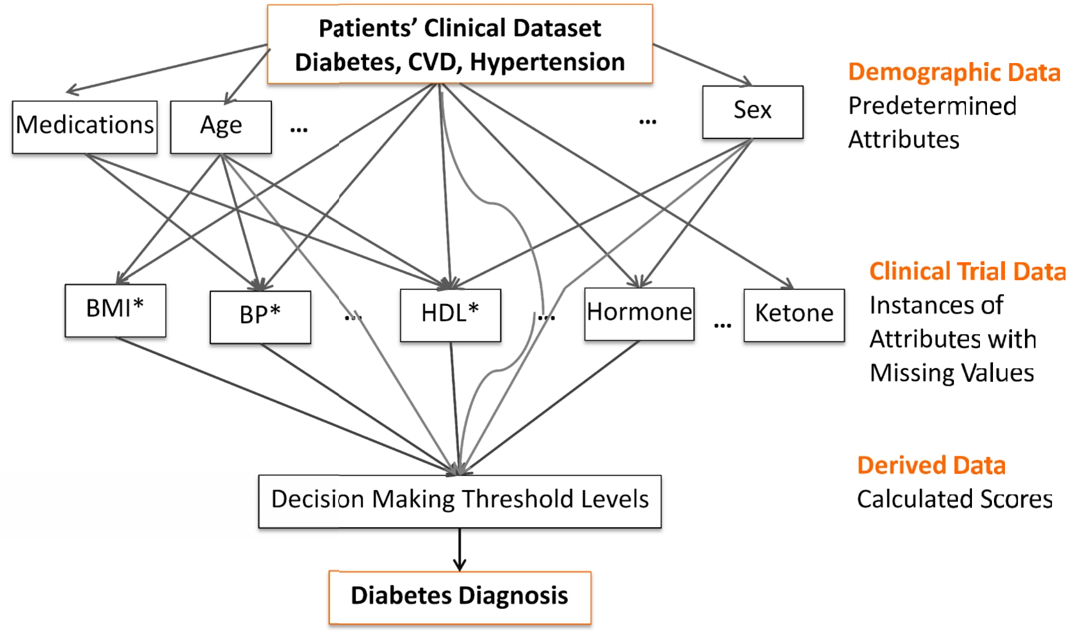

Fig. 3. Typical clinical dataset model

**BMI-Body Mass Index; BP-Blood Pressure; HDL- High Density Lipoprotein*

2.2 Classifiers

2.2.1 Decision tree

The tree induction algorithms iD3, C4.5, C5 advanced by [18] partition data in a hierarchical manner in the form of a tree where branches are represented by values of a particular selected attribute. An optimal tree is formed by selecting a feature at each node that gives the highest Information Gain (IG). Highly mixed partitions are nominated for further subdivision. In this work, we pursue a simpler version of DT where no model of data is learned but an appropriate leaf is directly accessed by mining through the training set. It is appropriate to call this DT variety - Decision Branch (DB).

2.2.2 Nearest neighbor

The NN algorithm used in this study is an adaptation of the well-known k-NN technique [41] for nominal data. It uses the Hamming loss for a distance function in the pseudo-space of data attributes. The loss, which normally counts dissidence of attribute values of two compared instances, is weighted by Information Gain (IG) to make the space metric conform better to the data [42,43].

2.2.3 Naïve Bayes

This is a classical method and widely used on nominal data despite the limiting assumption of attribute conditional independence, given a class [44].

2.2.4 Classifying with MVs

The NB classifier is easily adaptable for data with MVs since any test instance can be evaluated with some attributes withheld. However, this property also holds for NN and DB methods. This feature of classification algorithm implementation is the short memory, or lazy learning mode, whereby the learning is carried out anew for every test instance. Indeed, this mode offers a general method of obviating MVs when classifying any new data.

2.3 Filling-in the Blanks

For a discrete or categorical variable with MVs, some dummy value can be assigned to denote the missing value (e.g. “MV”). However, such an assignment of a dummy value outside the attribute’s domain cannot be carried out as elegantly with numerical features. Statisticians are known to have resorted to the insertion of dummy data (e.g. “999.999”) that invariably distorts distance metrics underpinning classifiers. The distortion of results when using distance metrics then provides an additional rationale to discretize continuous attributes in mixed attribute type domains. If MVs are sparsely distributed, they can be ignored during the learning and generalization steps of the algorithm. However, this is not possible for the data in this study because of the highly expressed attribute patterns of MV inundation.

Some condensation of incomplete data can also be achieved by discarding instances or attributes replete with MVs. However this cannot be a general method as only the training set can be dealt with in such a manner. Therefore incomplete information dismissal is proposed as a pre-processing step in the next section and is the first part of our approach for MV handling.

2.3.1 Incomplete information dismissal

Data layers, namely instances or attributes, are eliminated one by one, whichever currently conveys the least information, until a predetermined data reduction effect is achieved. Attribute information is calculated using IG, and is associated with any value unless it is missing. An attribute may be dismissed, even though it may have no MVs, which doubles as a feature selection step [11].

A data completion algorithm as proposed in this research utilizes the ability of classification algorithms to classify data without referencing MVs. The knowledge of class of an instance being dealt with provides a clue to MV substitution from applicable ranges. However, it is not required to label all of the classifier training dataset. Data completion forms the second part of the proposed technique.

2.3.2 Data completion

Our algorithm performs completion and classification of test instances at the same time by substituting the attribute modes for MVs from an applicable sample of training data, which a classifier accesses to make the prediction. It is necessarily an iterative process of tuning instances in turn to the rest of the set that requires at least one valued instance per attribute per class. MVs are substituted from the subset with the same class label as the test instance within the extracted sample, whether the predicted class is the same or not. In the case where the test instance is not labeled, the respective classification algorithm is used to set the label. The iteration ends when the introduced values stop changing or a limit for the number of cycles is reached. The goal is to achieve the highest possible classification accuracy, which the training set assumes, even though the instances may be rare.

2.4 Assembling an Appropriate Dataset

Despite the preponderance of patient data collected in digital forms, assembling an appropriate dataset to explore efficient diabetes management methods is inherently difficult because of a paradox in the formulation of the analytics exercise: on one hand, one cannot readily frame a question of interest without knowing what data is available, and on the other, one cannot identify what data is required without knowing the question of interest. The paradox presents itself in data mining process models such as CRoss Industry Standard Process for Data Mining (CRSIP-DM) [45], where the need to identify business objectives is advanced as a first step so that appropriate questions can be formulated to guide the selection of data. There are no clear ‘business’ objectives in diabetes mining, as the questions of interest are usually framed as high level questions such as “Identify risk factors for diabetes”, or “Predict blood glucose level” that cannot inform the selection of data, because variables that might be relevant are not known. In this study we advance the claim that the paradox inherent in assembling a dataset for diabetes mining necessitates the use of an extensive set of features that not only consists of features relevant only to diabetes but covers a range of other health and clinical indicators.

3. RESULTS AND DISCUSSION

3.1 Classifier Performance

The performance of Naive Bayesian (NB), Decision Branch (DB) and Nearest Neighbor (NN) classifiers in diagnosing diabetes, CVD and hypertension when applying the traditional model in Fig. 4 with MVs skipped, is compared to the accuracy after MV imputation using different strains of the proposed method in Fig. 5. The strains are denoted as NB, DB, and NN - same as the classifiers at the core of the principal method. The accuracy is calculated using leave-one-out cross-validation which is a conventional technique when a comparison is involved.

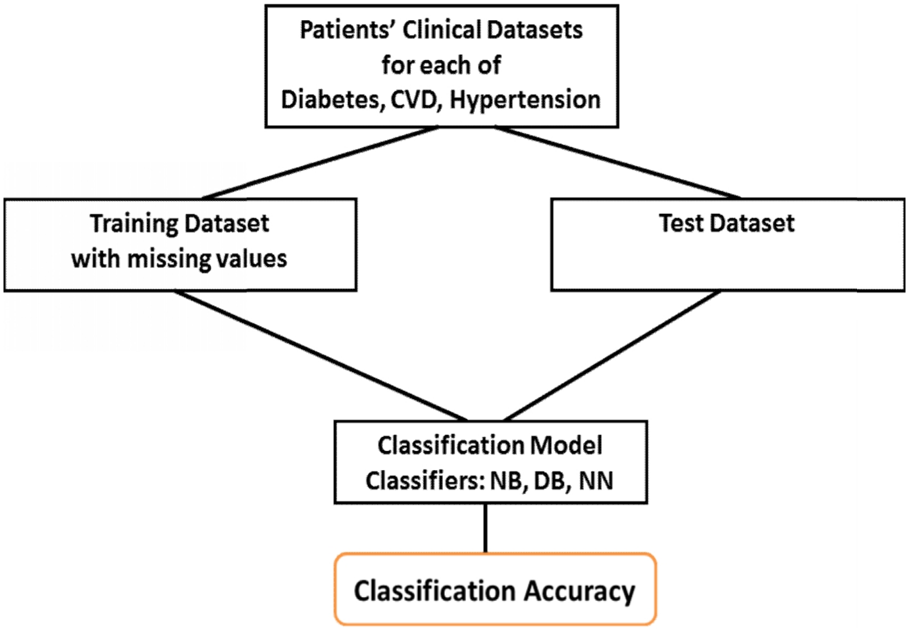
Patients' Clinical Datasets for each of Diabetes, CVD, Hypertension
Training Dataset with missing values
Test Dataset
Classification Model Classifiers: NB, DB, NN
Classification Accuracy

Fig. 4. Traditional classification model with missing values

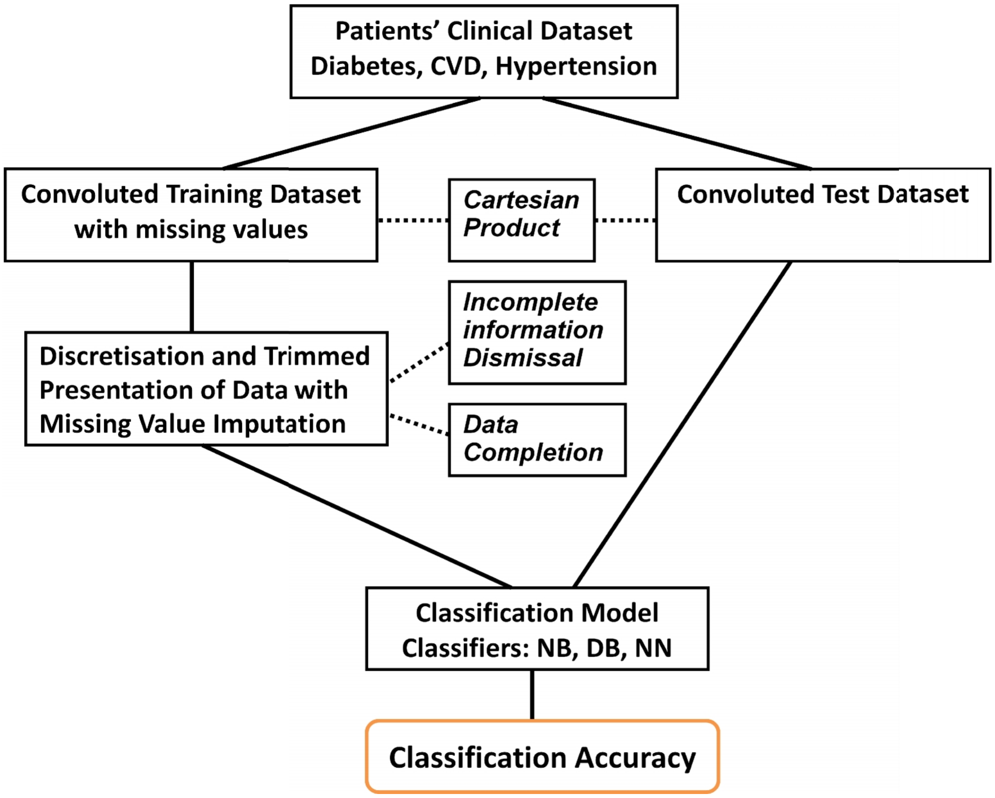
Patients' Clinical Dataset Diabetes, CVD, Hypertension
Convoluted Training Dataset with missing values
Cartesian Product
Convoluted Test Dataset
Incomplete information Dismissal
Discretisation and Trimmed Presentation of Data with Missing Value Imputation
Data Completion
Classification Model Classifiers: NB, DB, NN
Classification Accuracy

Fig. 5. Proposed classification model with missing value imputation

Fig. 5 provides an overview of the proposed classification model using our novel MV imputation techniques. The transformation of the training and test set involves convolution of class values for diabetes, hypertension and cardiovascular disease using *Cartesian Product*. The *Incomplete Information Dismissal* and *Data Completion*, after having any continuous features discretized, are performed prior to exposure of the dataset to the three classifiers. These three moments form the main novelty of our proposed model.

To evaluate the performance of our proposed *Incomplete Information Dismissal* method, we obtain a smaller dataset by reducing the amount of data by 50%. Fig. 6, used for illustration, provides the overall, any-class accuracy for both full and reduced datasets after performing data completion using the DB strain of the proposed method. It is evident from Fig. 6 that the reduction not only does not impact on the diagnostic ability, but also gives an improvement in some cases. At the same time, about 40% of MVs have been dealt with, without having to fill them. Different results pertain to different sub-problems featured within the DiScRi data.

To illustrate the performance of our proposed *Data Completion* method, the accuracy of the three classifiers when sidestepping MVs and after entering them using the DB strain of the principal method is compared in Fig. 7. For example, with DB substituted MVs, the NN classifier achieved 97% accuracy for CVD diagnosis with the full dataset as compared to 98% with the reduced dataset. High accuracies were also achieved for the diagnosis of diabetes (94% to 95%), with some exception for hypertension (90% to 89%), which had a minor decrease, using this combination. Similar results were obtained for NB and NN based completion methods for each of the three classifiers.

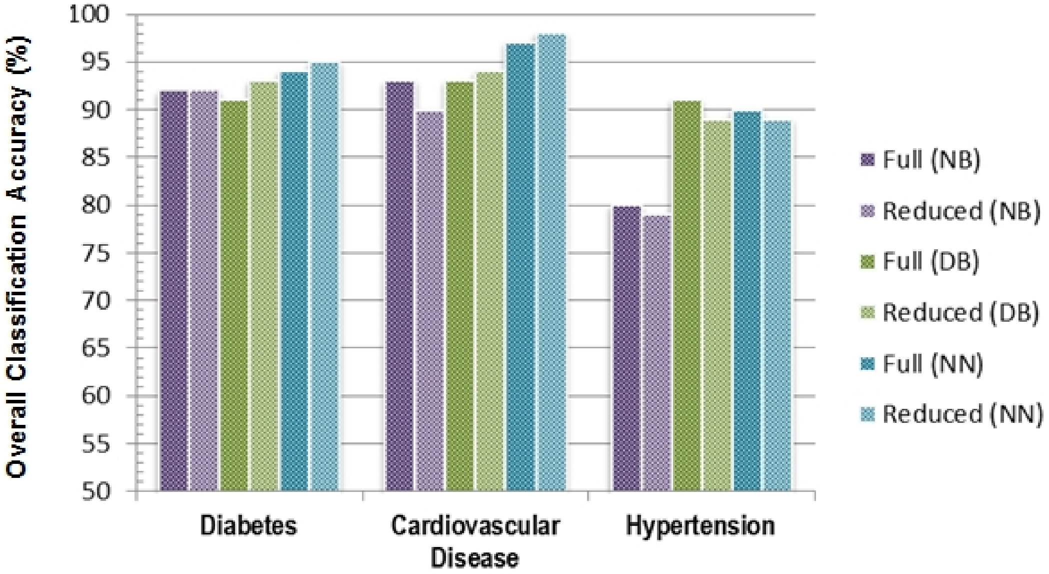

Fig. 6. Overall classification accuracy (%) on full and reduced datasets - by different classifiers (NB, DB, NN) after substituting MVs with Decision Branch inspired surrogates

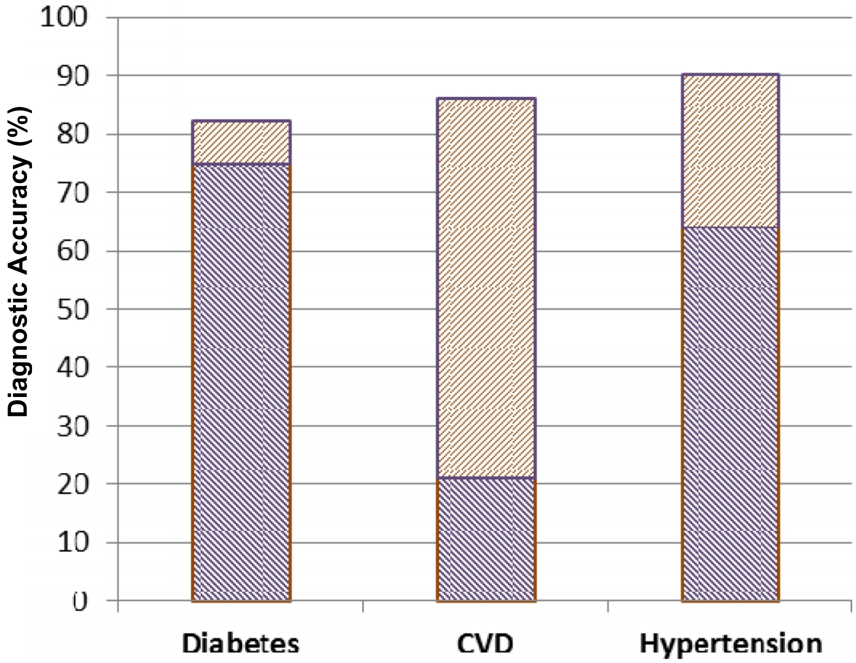

Fig. 7. Improvement of diagnostic accuracy after MV submission after applying the cartesian product

Dark portion: Class 1 diagnostic accuracy with MVs skipped; Light portion: Accuracy gain after Data Completion using the DB strain of the proposed method. The classifier is NN

3.2 Benchmarking

We staged a super problem where classes are defined by *Cartesian Product* of class values for diabetes, CVD and hypertension: this is a novelty of our approach that complements our unique imputation techniques of *Incomplete Information Dismissal* and *Data Completion*. For example, an instance may have the new attribute value as YNY (Yes for diabetes, No for CVD status and Yes for hypertension). This ‘additional’ class is used to take advantage of the nearest points formulation to cater for incomplete information dismissal and data completion. The advantage is that MVs on other attributes are imputed in a way that exalts the new triple attribute.

The improvements in the diagnostic accuracy of our algorithm are benchmarked using the three classifiers. The results of NN accuracy for full data of Class 1 is shown in Fig. 7 for both before and after data completion by the DB based method.

3.3 Discussion

Overall, from the performance results it is evident that the diagnostic predictions are better after data completion. While each classification method has its own merits and demerits, a particular advantage of our data completion procedure is that it imparts better estimation of any involved probabilities effectively extending the feature-set.

The classifiers sidestepping MVs achieve low accuracy in the diagnosis of diabetes, cardiovascular disease as well as hypertension. Also, when MVs are ignored and replaced with dummy data, the results are similar. This is probably due to the attribute rather than instance pattern of inundation of data with MVs, as shared by the entire dataset. Hence, we can infer that the affected attributes with MVs are predominantly irrelevant.

The performance of the NB based MV submission method in the current work stands out. The algorithm not only completes the iterations much faster but also rather accurate. The breadth of mode selection in the algorithm causes high contraction of data distribution in the pseudo-space of attributes, that is, classes become identified by the distinct feature modes. The DB guided MV submission performed slower and somewhat less accurate over the NB counterpart, although any possible improvement may have been restricted due to the data. The NN based data completion method shows the least improvement, although the accuracy is still acceptable. This is quite contrary to the usual high performance of the NN classification method and has to do with choices available for MV selection that are more limited than in completion with DB.

In medical diagnostic problems, the attribute pattern of MV inundation is often prevalent. Reducing the data set by 50% leaves all instances intact, though this is not the case with the number of attributes. Our algorithm tries to withhold attributes that are either less informative or with many MVs. In either case, trimming by half appears to be safe for all diabetes linked problems as the results are almost unchanged. Nonetheless, if too many features are removed the accuracy is expected to drop. Also, even though individual features are ranked by overall information gain, and attribute reduction occurs only for those that are least informative and found at the end of the list, improvement in accuracy may not be always guaranteed. However, the computational aspect of performance improves dramatically when such superfluous attributes are discarded.

While the results obtained are good, filling many missing entries could be misleading. The data is too flexible, and the high accuracy may be unwarranted. In this regard, the more classes are embraced by a problem, the better. However, here there are only two for any featured condition. There are no other constraints to narrow the range of MVs further. By linking the diabetes, cardiovascular and hypertension classes, additional clues can be arrived at. The three problems together interpret the same data, and so we expect the substituted values to be the same. Hence, we have used a convolution of the three to stage a super problem where classes are defined by the Cartesian product of class values of the underlying problems. Cartesian attribute products in classification were made popular by [46]. Here, each combination of the values corresponds to a class in the super problem. This forms the main novelty of our approach of MV imputation as compared to existing ones in the literature. In our model comprising the three problems with two classes, the Cartesian product generates eight classes same in all problems, thereby imposing much tougher constraints on MV ranges. This approach is innovative and provides an advancement towards an integrated and holistic management of diabetes.

4. CONCLUSION

Large clinical data collected from patients often include missing values or incomplete data, which poses a major problem for an integrated diagnosis. Data mining techniques dealing with such incomplete data problems in clinical trials become popular. This work demonstrated a considerable improvement in diagnostic accuracy after adopting the proposed methodology for missing value submission. The data completion methodology can be applied to data of any type via discretization of continuous attributes. The advancement is significant because quite high predictive accuracies, as measured with leave-one-out cross-validation resampling, appear to be achievable even when datasets contain substantial missing data.

Future work could involve the development of decision rules derived through the learning mechanism of our novel missing value handling techniques of *Incomplete Information Dismissal* and *Data Completion* in order to arrive at a decision tree model for new patients with clinical data. This would provide an enhanced decision support system for an integrated diabetes management, despite missing values in the patient's clinical data.

COMPETING INTERESTS

The authors declare that they have no competing interests.

REFERENCES

1. Devendra D, Liu E, Eisenbarth GS. Type 1 diabetes: recent developments. British Medical Journal (BMJ). 2004;328(7442):750-754.
2. Lee M, Saver JL, Hong KS, Song S, Chang KH, Ovbiagele B. Effect of pre-diabetes on future risk of stroke: Meta-analysis. British Medical Journal (BMJ). 2012;344(e3564):1-11.
3. Fagherazzi G, Vilier A, Bonnet F, Lajous M, Balkau B, Boutron-Ruault MC, Clavel-Chapelon F. Dietary acid load and risk of type 2 diabetes: The E3N-EPIC cohort study. Diabetologia. 2014;57(2):313-320.
4. Saydah SH, Fradkin J, Cowie CC. Poor control of risk factors for vascular disease among adults with previously diagnosed diabetes. Journal of the American Medical Association (JAMA). 2004;291(3):335-342.
5. Thunander M, Petersson C, Jonzon K, Fornander J, Ossiansson B, Torn C, Edvardsson S, Landin-olsson M. Incidence of type 1 and type 2 diabetes in adults and children in Kronoberg, Sweden. Diabetes Research Clinical Practice. 2008;82(2):247-255.
6. Lammi N, Blomstedt PA, Moltchanova E, Eriksson JG, Tuomilehto J, Karvonen M. Marked temporal increase in the incidence of type 1and type 2 diabetes among young adults in Finland. Diabetologia. 2008;51(5):897-899.
7. WHO. Diabetes, Fact sheet No. 312. World Health Organization Department of Noncommunicable Disease Surveillance: Geneva; 2012.
8. American Diabetes Association. (Standards of medical care in diabetes-2013. Diabetes Care. 2013;36(1):11-66.
9. Stranieri A, Zeleznikow J. Knowledge discovery from legal databases, Springer: New York; 2005.
10. Latkowski R, Mikołajczyk M. Data decomposition and decision rule joining for classification of data with missing values. In Transactions on Rough Sets I, LNCS 3100:299-320. Springer: Berlin Heidelberg; 2004.
11. Bagirov A, Yatsko A, Stranieri A, Jelinek HF. Feature selection using misclassification counts. *Proceedings of the* 9-th Australasian Data Mining Conference (AusDM 2011), Conferences in Research and Practice in Information Technology (CRPIT). 2011;121:51-62.
12. Meng X, Huan, Y, Rao D, Zhang Q, Liu Q. Comparison of three data mining models for predicting diabetes or prediabetes by risk factors. The Kaohsiung Journal of Medical Sciences. 2013;29(2):93-99.

13. DeFronzo RA. *International Textbook of Diabetes Mellitus*. 3rd ed. Chichester, West Sussex, Hoboken, John Wiley: New Jersey; 2004.
14. Kim SH. Measurement of insulin action: A tribute to Sir Harold Himsworth. Diabetic Medicine. 2011;28(12):1487-1493.
15. CDC. National diabetes fact sheet: National estimates and general information on diabetes and pre-diabetes in the United States 2011, U.S. Department of Health and Human Services, Centers for Disease Control and Prevention: Atlanta; 2011.
16. Triplitt CL. Examining the mechanisms of glucose regulation. American Journal of Managed Care. 2012;18(1):4-10.
17. Roumie CL, Hung AM, Greevy RA, Grijalva CG, Liu X, Murff HJ, Elasy T, Griffin MR. Comparative effectiveness of sulfonylurea and metformin monotherapy on cardiovascular events in type 2 diabetes mellitus: A cohort study. Annals of Internal Medicine. 2012;157(9):601-610.
18. Quinlan JR. Induction of decision trees. Machine Learning. 1986;1(1):81-106.
19. Kelarev AV, Abawajy J, Stranieri A, Jelinek HF. Empirical investigation of decision tree ensembles for monitoring cardiac complications of diabetes. International Journal of Data Warehousing and Mining. 2013;9(4):1-18.
20. Quinn A, Stranieri A, Yearwood J. AWSum - applying data mining in a health care scenario. International Conference on Intelligent Sensors, Sensor Networks and Information Processing. 2008:291-296.
21. Tseng SM, Wang KH, Lee CI. A preprocessing method to deal with missing values by integrating clustering and regression techniques. Applied Artificial Intelligence. 2003;17:535-544.
22. Junninen H, Niska H, Tuppurainen K, Ruuskanen J, Kolehminen M. Methods for imputation of missing values in air quality data sets. Atmospheric Environment. 2004;38:2895-2807.
23. Zhang S, Zhang J, Zhu X, Qin Y, Zhang C. Missing value imputation based on data clustering. Transactions on Computational Science, Springer, LNCS 4750. 2008;1:128-138.
24. Wang Q, Rao JNK. Empirical likelihood-based inference in linear models with missing data. Journal of Statistics. 2002;29:563-576.
25. Schneider T. Analysis of incomplete climate data: Estimation of mean values and covariance matrices and imputation of missing values. Journal of Climate. 2001;14(5):853-871.
26. Little RJA, Rubin DB. Statistical analysis with missing data. 2nd ed., Wiley: UK; 2002.
27. Farhangfar A, Kurgan L, Dy J. Impact of imputation of missing values on classification error for discrete data. Pattern Recognition. 2008;41(12):3692-3705.
28. Rahman G, Islam Z. A decision tree-based missing value imputation technique for data pre-processing. Proceedings of the 9-th Australasian Data Mining Conference (AusDM 2011), Conferences in Research and Practice in Information Technology (CRPIT). 2011;121:41-50.
29. Van Buuren S. Multiple imputation of discrete and continuous data by fully conditional specification. Statistical Methods in Medical Research. 2007;16(3):219–242.
30. Janssen KJ, Donders AR, Harrell FE Jr, Vergouwe Y, Chen Q, Grobbee DE, Moons KG. Missing covariate data in medical research: to impute is better than to ignore. Journal of Clinical Epidemiology. 2010;63(7):721–727.
31. Carpenter J, Kenward, M. Multiple Imputation and its Application, Wiley: UK; 2013.
32. Nevalainen J, Kenward MG, Virtanen SM. Missing values in longitudinal dietary data: a multiple imputation approach based on a fully conditional specification. Statistics in Medicine. 2009;28(29):3657-3669.

33. Marston L, Carpenter JR, Walters KR, Morris RW, Nazareth I, Petersen I. Issues in multiple imputation of missing data for large general practice clinical databases. Pharmacoepidemiology Drug Safety. 2010;19(6):618-626.
34. Sterne JAC, White IR, Carlin JB, Spratt M, Royston P, Kenward MG, Wood AM, Carpenter, JR. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. British Medical Journal (BMJ). 2009;338(6):157-160.
35. Marinov M, Mosa ASM, Yoo I, Boren SA. Data-mining technologies for diabetes: A systematic review. Journal of Diabetes Science and Technology. 2011;5(6):1549-1556.
36. Jelinek HF, Wilding C, Tinely P. An innovative multi-disciplinary diabetes complications screening program in a rural community: A description and preliminary results of the screening. Australian Journal of Primary Health. 2006;12(1):14-20.
37. Cornforth DJ, Jelinek HF, Teich MC, Lowen SB. Wrapper subset evaluation facilitates the automated detection of diabetes from heart rate variability measures. Proceedings of the International Conference on Computational Intelligence for Modelling Control and Automation (CIMCA'2004). 2004;446-455.
38. Abawajy J, Kelarev A, Chowdhury M, Stranieri A, Jelinek HF. Predicting cardiac autonomic neuropathy category for diabetic data with missing values. Computers in Biology and Medicine*.* 2013;43(10):1328-1333.
39. Kelarev AV, Stranieri A, Yearwood J, Jelinek HF. A comparison of machine learning algorithms for multilabel classification of CAN. Advances in Computer Science and Engineering. 2012;9(1):1-4.
40. Yang Y, Webb GI. Discretization for naive-bayes learning: Managing discretization bias and variance. Machine Learning. 2009;74(1):39-74.
41. Keller JM, Gray MR, Givens JA. A fuzzy k-nearest neighbor algorithm. IEEE Transactions on Systems, Man and Cybernetics*.* 1985;15(4):580-585.
42. Daelemans W, Van-Den-Bosch A. Generalization performance of backpropagation learning on a syllabification task. Proceedings of TWLT3 - the Third Twente Workshop on Language Technology*):*27-37. Enschede, Morgan Kaufmann: Netherlands; 1992.
43. Wettschereck D, Aha DW, Mohri T. A review and empirical evaluation of feature weighting methods for a class of lazy learning algorithms. Artificial Intelligence Review*.* 1997;11:273-314.
44. Domingo SB, Pazzani M. Beyond independence: Conditions for the optimality of the simple Bayesian classifier. Proceedings of the Thirteenth International Conference on Machine Learning). Morgan Kaufmann: Netherlands. 1996;105-112.
45. Wirth R, Hipp J. CRISP-DM: Towards a standard process model for data mining. In Proceedings of the 4th International Conference on the Practical Applications of Knowledge Discovery and Data Mining. 2000;29-39.
46. Pazzani MJ. Constructive induction of Cartesian product attributes. In Liu H, Motoda H. (Ed.), Feature extraction, construction and selection: A data mining perspective). Kluwer Academic Publishers: Netherlands. 1998;341-354.

Peer-review history:
The peer review history for this paper can be accessed here:
http://www.sciencedomain.org/review-history.php?iid=670&id=5&aid=6128