BIOINFORMATION
discovery at the interface of physical and biological sciences

OPEN ACCESS GOLD

BIOMEDICAL INFORMATICS

www.bioinformation.net
Volume 15(12)

Research Article

Semi supervised data mining model for the prognosis of pre-diabetic conditions in type 2 Diabetes Mellitus

A. Sumathi[1*] & S. Meganathan[2]

[1]Department of Computer Science Engineering, SASTRA Deemed To be University, Srinivasa Ramanujan Centre (SRC), Kumbakonam, Tamil Nadu, India; [2]Department of Computer Science Engineering, SASTRA Deemed To be University, Srinivasa Ramanujan Centre (**SRC**), Kumbakonam, Tamil Nadu, India; E-mail – sumathi@src.sastra.edu, meganathan@src.sastra.edu; *Corresponding author

Received December 16, 2019; Revised December 26, 2019; Accepted December 29, 2019; Published December 31, 2019

DOI: 10.6026/97320630015875

Abstract:

Diabetic Mellitus is the leading disease in the world irrespective of age and geographical location. It is estimated that 43% of the overall population is affected by the disease. The reasons for the disease include inappropriate diet lifestyle with allied symptoms like obesity. Therefore, the prognosis and diagnosis of the disease are important for adequate combat and care. The prognosis related known symptoms of the disease include incontinence (inability to control urination) and frequent fatigue. Moreover, early prediction of the disease plays an important role in the prognosis of other associated conditions such as heart failure leading to chronic illness. Hence, it is of interest to describe a data mining based prediction model using known features (derived from epidemiological data collected from the public hospital using routine tests) to help in the prognosis of the disease. We used data pre-processing techniques for handling missing values and dimensionality reduction models to improve data quality. The Minimum Description Length principle (MDL) model for discretization (replacing a continuum with a finite set of points) is used to reduce high-level dimensionality of the dataset, which enabled to categorize the dataset into small groups in ordered intervals. Thus, we describe a semi-supervised learning technique (identifies promising attributes using clustering and classification methods) by combining data mining techniques for reasonable accuracy having adequate sensitivity and specificity for further discussion, cross-validation, revaluation, and application. Early prediction of the disease with improved accuracy by analysing specificity ranges in blood pressure and glucose levels will be useful to combat Diabetes Mellitus.

Keywords: Diabetes, semi supervised learning, epidemiological data, prognosis, classification, clustering

Background:

The world has become health conscious. People are interested in understanding the root cause of diseases. Most diseases are caused in two major ways. The first cause is uncontrolled food intake and other cause is due to alcohol and tobacco consumption. The disproportionate intake of unhealthy food is linked with diseases like cancer, obesity, etc. However, inadequate intake of food causes anemia, mineral deficiency, mal nutrition and other diseases. Diabetes Mellitus **[1]** is one of the life-threatening diseases, which causes due to dysfunction of the pancreas due to insufficient secretion of insulin irrespective of age. Diabetic Mellitus is a silent killer **[2]**. The known subsidiary diseases of diabetes are retinopathy, heart disease, and chronic kidney disease. Glucose test before and after food intake is common **[3, 4]**. Data Mining **[5]** is a powerful technique used to identify the hidden patterns from data repositories. The prognosis of pre diabetes symptoms using known epidemiological data is useful in treatment. The standard data mining algorithms are used to identify, analyze the hidden data patterns and collect relationships in known stored data. The building of interconnected relations in known clinical data shows different combinations of symptoms towards combat and care. Therefore, it of interest to use machine learning **[5,6]** and especially Semi-supervised learning [7,8] exploiting known classification and clustering techniques to help in the prognosis diabetes (**Figure 1**).

BIOINFORMATION
discovery at the interface of physical and biological sciences

Data Mining in Biological domain:
Data Mining is one of the oldest and standard techniques for handling biological data in different forms **[8,9].** The inherent nature of the complexity of biological data requires a collection of techniques to analyze the data from large datasets. The information retrieval is possible using data mining with contextual results relating multiple dimensions. The use of Machine learning for data mining in large data is becoming common. Further, the use of the semi-supervised learning **[9]** in analysing both labelled and unlabeled data is getting familiar. Thus, the use of machine learning especially semi-supervised learning approaches in biological data mining finds immense application in biological knowledge discovery **[10]**.

Methodology:
Data mining techniques are used for medical applications. The single approach always gives results for a particular approach with less optimal output. A combination of approaches is needed to handle a different set of attributes with interrelationships. Data pre-processing **[11]** is primary in data mining. The use of MDL (Minimum Description Length) in diabetes linked data mining is known. Our interest is to use machine learning especially semi-supervised learning approaches in diabetes linked data mining to glean useful information related to the disease.

Preprocessing:
Data preprocessing **[12]** is a necessary to remove noise to improve accuracy. The calculation of the average mean value in a given dataset is primary in this context. The formula used for mean is given below:

The Mean (x) = (sum) of all the values $\sum(x)$ / the number of values (n).

$mean(x) = \sum(x)/n$

The x represents the inputs of patient attribute values.

Discretization:
Biological data is multidimensional. Hence, minimization of its dimensionality is the first step for categorization. The discretization **[13]** minimizes the distance between related values of attributes by ordering it. The ordered set is well organized with minimal and maximum intervals. It reduces the large chunks of numeric values into a group of well-organized values. The discretization contains many techniques. MDL is a suitable method for identifying the most promising attribute.

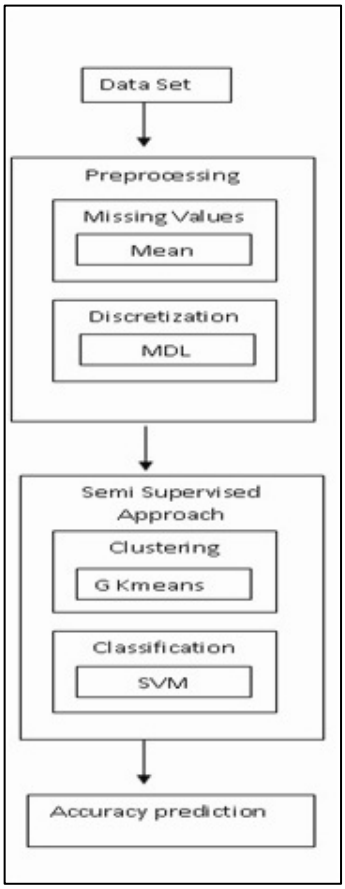


Figure 1: Semi supervised learning model for predicting pre-diabetes occurrences

Minimum Description Length (MDL):
MDL is based on information theory and entropy using data representation to find regular patterns in data **[14].** MDL helps to filter the most relevant values in the normal values. We used the following entropy formula in MDL discretization.

$$Entropy(D1) = -\Sigma^{m}_{i=1} p_i \log_2 p_i$$

Clustering and Genetic *k-means* algorithm:
The Clustering technique is a simple and powerful technique that is used to categorize the unlabelled data into groups called clusters. Similar and relevant data points are organized into groups during clustering. It is an iterative technique, which uses updated center

BIOINFORMATION
discovery at the interface of physical and biological sciences

data point in each iteration. The limitation in clustering is its local maximum. The group needs a generalized scope to increase its scope in a cluster to achieve better scope. Global maximum is needed for generalized applications. The general k-means clustering algorithm with a genetic approach is combined to form a genetic *k* means algorithm to predict pre diabetic conditions. The genetic *k* means **[15]** is an evolutionary approach that gives optimum solutions for complex problems to achieve local maximum with good performance in a diabetic dataset (**Figure 2**).

The fitness function is applied for clustering is given below

$$P(Si) = F(Si)/\Sigma^{N}_{j=1}F(Sj)$$

Classification using Support Vector Machine (SVM):
SVM helps to distinguish data classes and concepts. Classification **[16, 17]** helps in identifying known data from unknown resources. The known data is called training data and the unknown data is called test data. In a semi-supervised learning approach, the Diabetic data set attributes and ranges are categorized using the clustering genetic k means method. Therefore, the set of categories (subpopulations) and new observation belongs to the basis of a training set of data containing observations and whose categories are known. The SVM **[17]** is a training algorithm, which builds a model that assigns new examples to one category, by making it a non-probabilistic binary linear classifier. The SVM performs well compared with other statistical or machine learning methods, especially with biological datasets. Kernel methods like the SVM can easily handle non-vector inputs unlike most machine learning methods. It evaluates prognostic models by separating the initial information set as a coaching data. Thus, the SVM classification rule provides higher accuracy in the result (**Figure 3**).

Dataset:
The dataset consists of 2106 person records collected from a Government Hospital, Kumbakonam, Tamil Nadu, India. The record consists of basic health checkup of an individual with 11 attributes that are all numeric. The attribute contains the habitual activities of a person, blood pressure, height, weight, BMI, blood glucose level, age, gender etc.

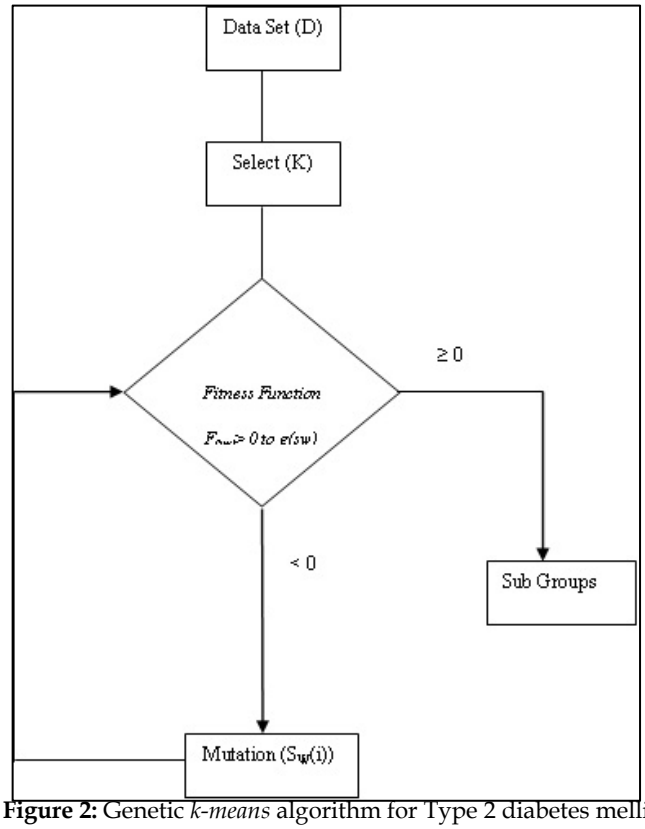


Figure 2: Genetic *k-means* algorithm for Type 2 diabetes mellitus prediction

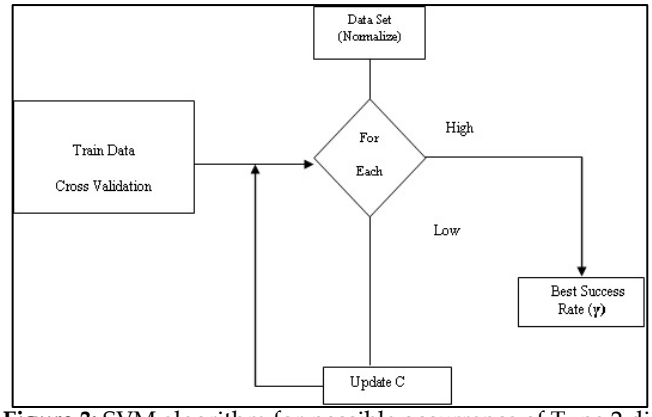
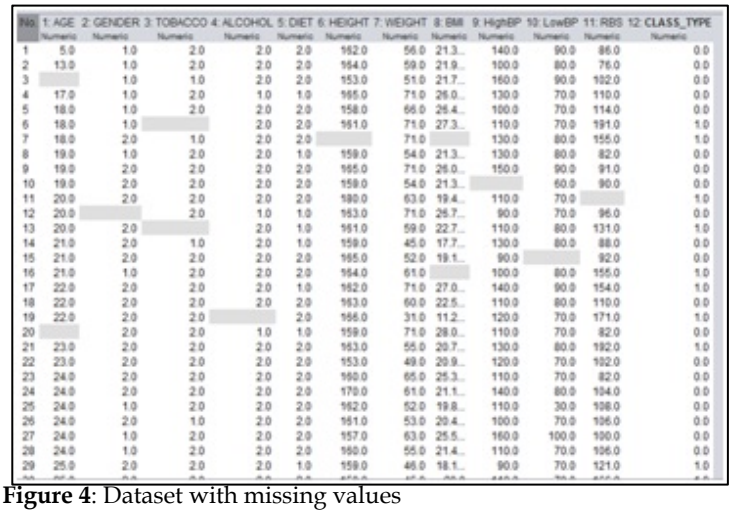


Figure 3: SVM algorithm for possible occurrence of Type 2 diabetes mellitus

BIOINFORMATION
discovery at the interface of physical and biological sciences



No	1: AGE	2: GENDER	3: TOBACCO	4: ALCOHOL	5: DIET	6: HEIGHT	7: WEIGHT	8: BMI	9: HighBP	10: LowBP	11: RBS	12: CLASS_TYPE
	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric
1	5.0	1.0	2.0	2.0	2.0	162.0	56.0	21.3...	140.0	90.0	86.0	0.0
2	13.0	1.0	2.0	2.0	2.0	164.0	59.0	21.9...	100.0	80.0	76.0	0.0
3		1.0	1.0	2.0	2.0	153.0	51.0	21.7...	160.0	90.0	102.0	0.0
4	17.0	1.0	2.0	1.0	1.0	165.0	71.0	26.0...	130.0	70.0	110.0	0.0
5	18.0	1.0	2.0	2.0	2.0	158.0	66.0	26.4...	100.0	70.0	114.0	0.0
6	18.0	1.0		2.0	2.0	161.0	71.0	27.3...	110.0	70.0	191.0	1.0
7	18.0	2.0	1.0	2.0	2.0		71.0		130.0	80.0	155.0	1.0
8	19.0	1.0	2.0	2.0	1.0	159.0	54.0	21.3...	130.0	80.0	82.0	0.0
9	19.0	2.0	2.0	2.0	2.0	165.0	71.0	26.0...	150.0	90.0	91.0	0.0
10	19.0	2.0	2.0	2.0	2.0	159.0	54.0	21.3...		60.0	90.0	0.0
11	20.0	2.0	2.0	2.0	2.0	180.0	63.0	19.4...	110.0	70.0		1.0
12	20.0		2.0	1.0	1.0	163.0	71.0	26.7...	90.0	70.0	96.0	0.0
13	20.0	2.0		2.0	1.0	161.0	59.0	22.7...	110.0	80.0	131.0	1.0
14	21.0	2.0	1.0	2.0	1.0	159.0	45.0	17.7...	130.0	80.0	88.0	0.0
15	21.0	2.0	2.0	2.0	2.0	165.0	52.0	19.1...	90.0		92.0	0.0
16	21.0	1.0	2.0	2.0	2.0	164.0	61.0		100.0	80.0	155.0	1.0
17	22.0	2.0	2.0	2.0	1.0	162.0	71.0	27.0...	140.0	90.0	154.0	1.0
18	22.0	2.0	2.0	2.0	2.0	163.0	60.0	22.5...	110.0	80.0	110.0	0.0
19	22.0	2.0	2.0		2.0	166.0	31.0	11.2...	120.0	70.0	171.0	1.0
20		2.0	2.0	1.0	1.0	159.0	71.0	28.0...	110.0	70.0	82.0	0.0
21	23.0	2.0	2.0	2.0	2.0	163.0	55.0	20.7...	130.0	80.0	192.0	1.0
22	23.0	2.0	2.0	2.0	2.0	153.0	49.0	20.9...	120.0	70.0	102.0	0.0
23	24.0	2.0	2.0	2.0	2.0	160.0	65.0	25.3...	110.0	70.0	82.0	0.0
24	24.0	2.0	2.0	2.0	2.0	170.0	61.0	21.1...	140.0	80.0	104.0	0.0
25	24.0	1.0	2.0	2.0	2.0	162.0	52.0	19.8...	110.0	30.0	108.0	0.0
26	24.0	2.0	1.0	2.0	2.0	161.0	53.0	20.4...	100.0	70.0	106.0	0.0
27	24.0	1.0	2.0	2.0	2.0	157.0	63.0	25.5...	160.0	100.0	100.0	0.0
28	24.0	1.0	2.0	2.0	2.0	160.0	55.0	21.4...	110.0	70.0	106.0	0.0
29	25.0	2.0	2.0	2.0	1.0	159.0	46.0	18.1...	90.0	70.0	121.0	1.0

Figure 4: Dataset with missing values

Results and Discussion:

The described model consists of three steps: (a) Data pre-processing; (b) Clustering and (c) Classification. The preprocessed data is then clustered using the genetic k-means technique producing labeled categorized data. The labeled data is then classified into subgroups based on the attribute intervals by using the SVM technique. The possible diabetic occurrence promising attribute values are identified using sub categorization done by SVM classification. The analysis is done in the WEKA **[18]** tool where the real-time obtained data is given as input. The missing values in the given data set are removed by replacing the average mean of an attribute. The data is loaded into the WEKA tool and missing values are identified **(Figures 4 and 5).** The missing values are replaced using the mean average replacement technique during preprocessing. The mean minimizes the error ratio with improved accuracy. The values for attributes are ordered. The dimensionality of attributes is reduced in the category. The preprocessed data is then analyzed using the MDL technique (**Figure 6**). The output of MDL implementation gives the ordered values of attributes (**Figure 7**). The second step of the model involves the labeling of data using Genetic k-means clustering. The preprocessed data set using mean and MDL are given as input. The initial cluster *k* value is given as 3 and the data set is divided into three labeled clusters whereas it contains sub-clusters with the nearest ranges (**Figure 8**). The third step is classification to identify the normal values in the range. The SVM uses correlation and regression to build relationships among the given attributes. Therefore, the sub-categorization of labeled data enables to predict promising attribute values of Type 2 diabetes mellitus occurrences. The performance of the SVM classifier was evaluated using the confusion matrix. The classifier results are shown in **Figure 9**. Analysis shows results from the preprocessing implementation using mean values in MDL. The semi-supervised learning approach is applied as a combined technique of genetic k-means and SVM classifiers in this study. The unlabelled data is categorized using the genetic *k-means* algorithm with optimal subgroups as clusters. The data points in each attributed values are updated during iteration for every biomarker. Each attribute values are clustered as a subgroup with nearby data points. The SVM classifier is used to categorize the relevant attributes as sub-groups. The classifier relates the attribute values and separates the promising attribute values from the normal range attribute values. The promising attributes values are the possible values of the occurrence of the disease. The most significant value in each attribute is correlated with the next promising attribute value to improve the accuracy rate in diabetic prognosis.

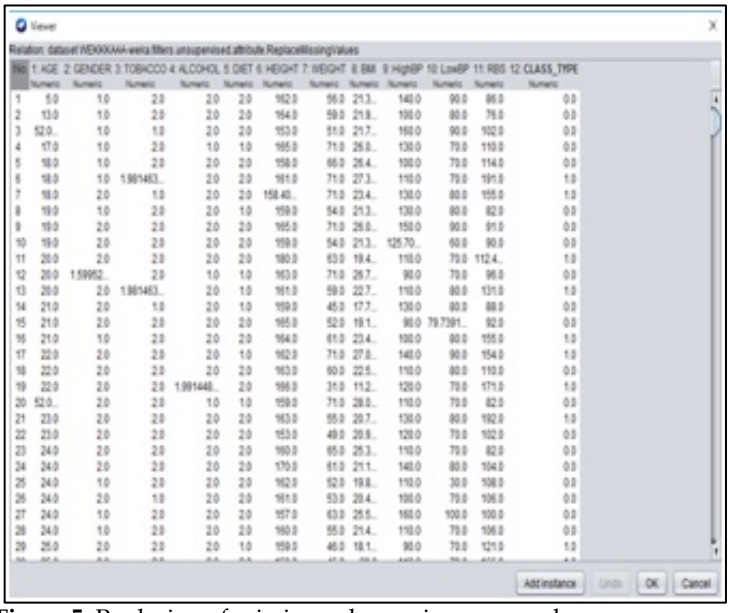


Figure 5: Replacing of missing values using mean values

BIOINFORMATION
discovery at the interface of physical and biological sciences

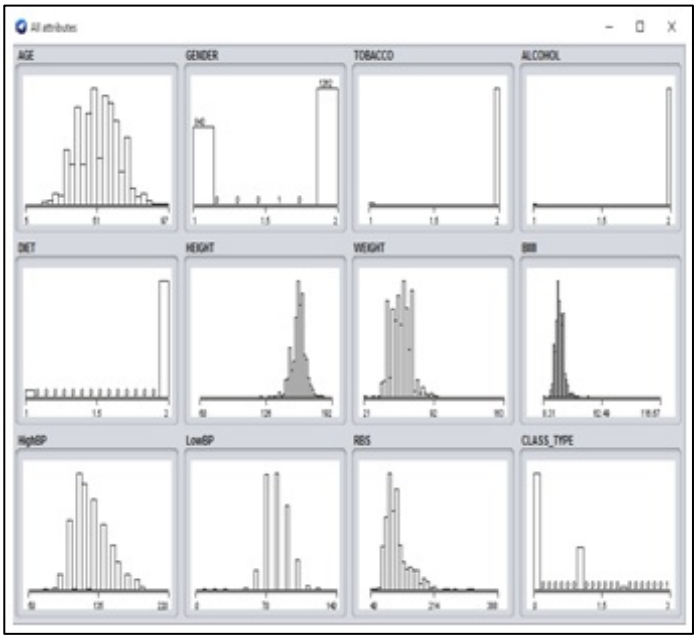


Figure 6: Dataset before discretization

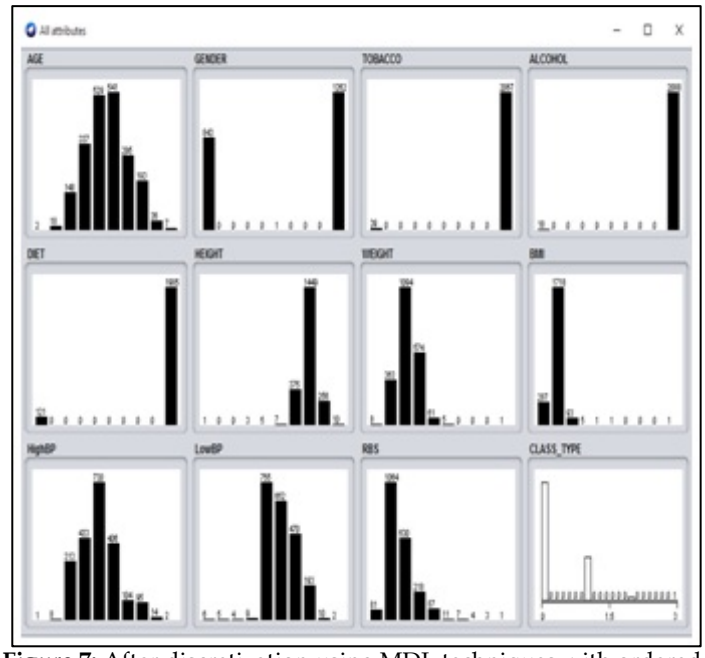


Figure 7: After discretization using MDL techniques with ordered attributes

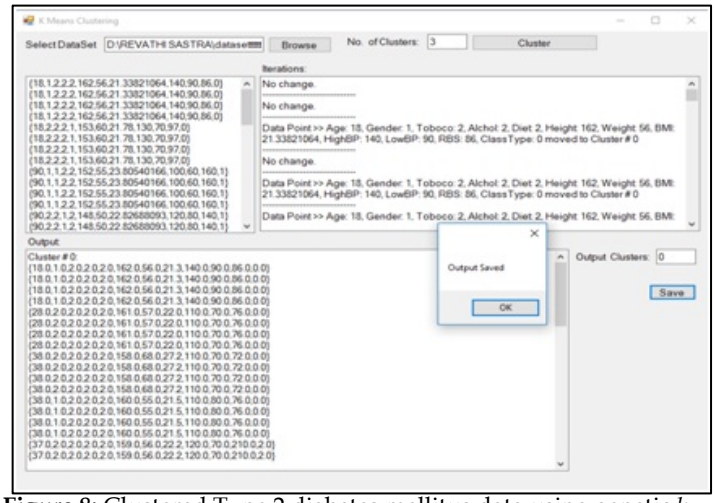


Figure 8: Clustered Type 2 diabetes mellitus data using genetic *k-means* algorithm

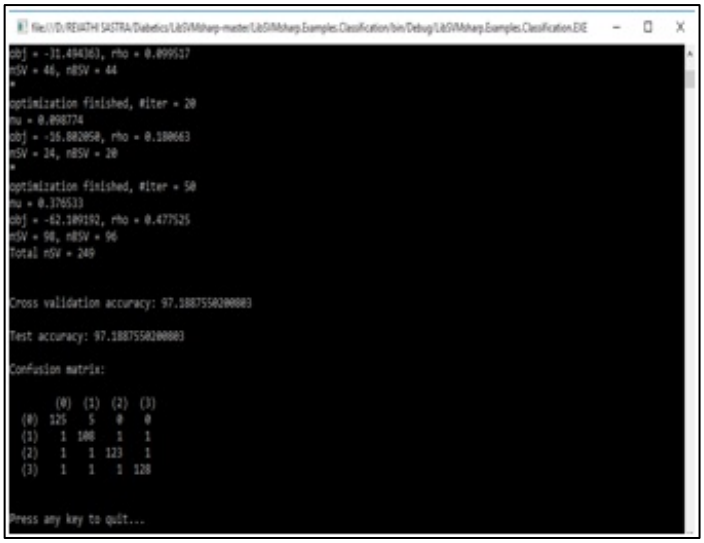


Figure 9: SVM based classification for Type 2 diabetes mellitus data

Conclusion:

We describe a semi-supervised knowledge discovery model by finding attributes using clustering and classification procedures in data mining for reasonable precision with sufficient accuracy for further debate, confirmation, revaluation and application in the prognosis of pre-diabetic conditions for Type 2 diabetes mellitus. We used the data collected from a public hospital for this analysis.

BIOINFORMATION
covery at the interface of physical and biological sciences

The dataset is used as training set for the model. The model identified the promising attributes using clustering and classification techniques. The data preprocessing techniques for handling missing values and dimensionality reductions are used to improve the quality of data. The MDL discretization technique reduced the high-level dimensionality of the dataset. This enabled to categorize the dataset into small groups in ordered intervals for improved accuracy.

Acknowledgments:
The open access charge for this article is partially sponsored by Biomedical Informatics (P) Ltd, India

Competing interests:
The authors declare no conflict of interest.

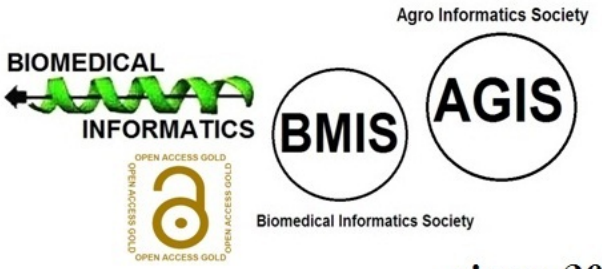
References:

[1] Tse SY, *BMC FamPract* 2018 **19**:199. [PMID: 30558542]
[2] Wanner C & Marx N *Diabetologia* 2018 **61**: 2134. [PMID: 30132035]
[3] Ikwuobe J *et al. Biogerontology* .2016 **17**: 511. [PMID: 26897532]
[4] Kirk IK *et al. Elife* 2019 [PMID: 31818369]
[5] Kraniotou C *et al. J Proteomics*. 2018.**188:**59 [PMID: 29518575]
[6] Kavakiotis I *et al. Comput Struct Biotechnol* 2017 **15**:104. [PMID: 28138367]
[7] Dagliati A *et al. J. Diabetes Sci Technol*.2018 **12**:2. [PMID: 28494618]
[8] Chapelle O, The MIT Press Cambridge, Massachusetts London, England (2006).
[9] Woldaregay AZ *et al. ArtifIntell Med.* 2019 **98**:109. [PMID: 31383477]
[10] Dong-giLee. *J. Expert Systems with Applications* .2018 **106**:15.
[11] Nie F *et al. IEEE Trans Image Process*. 2018 **27**:3. [PMID: 28945592]
[12] Bakariya B, *Advances in Intelligent Systems and Computing* Springer, India 2012 pp 202.
[13] Zhu M. *Communication Systems and Information Technology* 2011 pp 333
[14] Quinlan JR, *Kluwer Academic Publishers, Boston, Machine Learning* 1986 **1:** 81
[15] Paramveer S. Dhillon, *J. Machine Learning Research* 2011 **12:** 525.
[16] K. Krishna, *IEEE Transactions on Systems, Man, and Cybernetics*, Part B 1999 **29:**3
[17] Miroslav Marinov, *J Diabetes Sci Technol.* 2011 **5:** 1549. PMID: 22226277]
[18] Cui S, *Comput Methods Programs Biomed.* 2018 **166:**123. PMID: 30415712
[19] Smith TC, *Methods Mol Biol.* 2016 **1418:**353. [PMID: 27008023]

Edited by P Kangueane
Citation: Sumathi & Meganathan, Bioinformation 15(12): 875-881 (2019)

BIOINFORMATION
Discovery at the interface of physical and biological sciences

OPEN ACCESS GOLD



since 2005

BIOINFORMATION
Discovery at the interface of physical and biological sciences

indexed in