Application of Big Data on Diabetology: A Review

Lucy Sharma¹, Mohamed Sohel² & N.Divya³
¹,²,³Centurion University of Technology and Management. Department of CSE, Bhubaneswar, India

Abstract: Diabetes Mellitus has been the most common non-communicable disease since few recent centuries and according to reports, every eight out of ten adults are affected by this disease. After a certain stage of diabetes mellitus, a condition called diabetic retinopathy occurs which causes abrasions in the retina and progressively damages it. Therefore suitable diagnosis is very important to avoid the mutilation of other organs of the body. There is an ever-increasing rate of patients for the cure of this disease, consequently it generates large amount of data. A proper analysis of this data may provide new rays of hope in curing this disease. The biggest hurdle in this path is assortment, retrieval and analyzing large amount of data in fast and effective manner. This paper intends to provide an extensive review on the techniques of Hadoop's map reduce and Data mining algorithms like Naive Bayes, K-Nearest Neighbors, Vector Machine, ID3, C4.5, C5.0, CART for determining the patterns of diabetes of different individuals and their blood sugar status, so as to identify the unrevealed patterns related to the disease.

INTRODUCTION

Diabetes Mellitus is one of the non-communicable disease which has proved to be a major health hazard in this era. The number of patients which are being affected by this disease is increasing day by day and the bitter truth is that it can now even affect the unborn fetus. Because of this ever-increasing exposure to this disease large amount of data are produced which mainly comprises of Electrical Health Reports of patient’s data, clinical reports, data from social media and medicinal journals. These data form the big data for the medical industry. Through the analysis of this data, results can be predicted for understanding and improving the pathway of controlling the havoc of Diabetes Mellitus and creating a life time prospect by treatment at early stages. Predictive analysis is a system which integrates diverse techniques from data mining, statistics and game theory. This method incorporates the contemporary data and past data with statistical or other analytical model to determine or predict certain future events. This paper describes about the predictive analysis algorithm in Hadoop environment to predict the predominant diabetes type, impediments associated with it and the type of treatment to be employed.

The effect of diabetes on human body is such that if it crosses its particular range in the human blood, it can damage other organs of the body such as retina, where it leads to Diabetic Retinopathy, an acute disorder which injures the retina and is the major cause of blindness. The medical signs of Diabetic Retinopathy includes red lesions (RLs) which are the first obvious signs of DR. Subsequently the detection of RLs may prove to be a chief contribution in the screening tasks of DR.

The ever-increasing frequencies of diabetes, the number of images needed to examine retinal disorder is hefty. Thus, automatic analysis of retinal images in detection of Diabetic Retinopathy can help in diagnosis of the disease. The conversion of this huge amount of dense, contradictory, inaccurate and fragmented data into a structured one and then analyzing these data may provide breakpoints for curing add-ons of Diabetic Mellitus. The biggest hurdle in this step is that many valuable data are unseen which in turn is caused due to enormous amount of data. Applying Data Mining algorithms on medical data sets can bring an effective and revolutionary change in the process of diagnosing the disease and predicting the side effects according to the different stages of Diabetes Mellitus. This paper reviews about the effectiveness of different data mining algorithms based on the velocity, veracity and effectiveness of different algorithms like SVM, KNN, Naive Bayes, CART etc. In addition the paper also highlights how Diabetic retinopathy can be predicted using Big Data.

This paper is organized as follows: The chapter-1 describes about the different algorithms, its effectiveness and raises out the best technique used in analysis of data. Chapter-2 describes the predictive methodologies of Diabetic Data Analysis. Chapter-3 highlights on the prediction of
Diabetic Retinopathy one of the disease associated to Diabetic Mellitus.

1. DATA MINING ALGORITHMS IN CURING DIABETES

Data mining plays a crucial role in the diagnosis of diabetes because it can uncover the unknown facts related to the disease using enormous amount of medical data. Data mining aims to generate fresh realities which can be helpful by the medical experts, thus a practical model of curing the disease can be produced.

1.11 SVM (Support Vector Machine)

SVM is a binary classifier, that separates the data sets into two groups at a time, the objective of this algorithm is to find the best separating hyper plane which is also referred as decision boundary and by best we mean that the margin between the hyper plane and the data sets is greatest and the errors in classification comes out to be least. Through the linear and non-linear classification, the data is transformed to a higher dimension where it captures much more complex relationships between data points without having to perform difficult transformations.

1.12 KNN (K Nearest Neighbour)

KNN algorithm implies the technique in which a sample of data is segregated and other data is matched with the K sample and if the data with the properties same as the K sample are found it is kept in the same category. In other words, K of sample is identified for new sample and the label of the category which is more repetitive among these samples are recognized as the class of its result.

1.13 NAÏVE BAYES

Pattern recognition and classification problems is solved by the basic statistical method called Naïve Bayes. The problem for recognizing data between choices of different categories, probability distribution functions and values commensurated with them is used. Basically this algorithm works on various hypothetical and conditional possibilities.

1.14 ID3

ID3 decision tree uses statistical value a property and determines to what extent a property can separate ten different samples according to their categories. This tree classifies the samples to deep width so as to segregate and identify different patterns in the existing sample.

1.15 CART

CART algorithm divides the samples into two subsets in a way that records of each subset is more equal compared to the previous subsets. The division is repeated again and again until it faces the cease condition. It creates two branches per nodes. Each branch leads to another leaf node. Traversing a decision tree from root down, one row or value is ascribed to a case.

1.16 C4.5

C4.5 algorithm is similar to ID3 algorithm as it builds decision tree from the set of already classified dataset using the concept of information entropy. Entropy refers to the disorderness or uncertainty in the information. The training data set \( S = S_1, S_2, \ldots, S_n \) of samples. Each sample \( S_i \) consists of a p-dimensional vector \((x_{1i}, x_{2i}, \ldots, x_{pi})\) where \( x_i \) represents features of the sample as well as the class in which it falls. C4.5 chooses the attribute of the data that most effectually splits its set of samples augmented in one class at each node. The criterion for splitting is the difference between the entropy. The attribute with highest difference in entropy is chosen to make the decision.

1.2 PREVIOUSLY EVALUATED PERFORMANCE OF DM ALGORITHMS

By the proper analysis of the previous works on Data Mining Algorithm the efficiency of the algorithm is found out. The performance was evaluated by using 768 records of diabetic patients from PIMA dataset with eight important properties. The detection model used for evaluation are Precession, Recall, Measure and Accuracy.

\[
\text{Precession} = \frac{TP}{TP+FP} \\
\text{Recall} = \frac{TP}{TP+FN} \\
F\text{-measure} = \frac{2 \times \text{Precession} \times \text{Recall}}{\text{Precession} + \text{Recall}} \\
\text{Accuracy} = \frac{(TP+TN)/(TP+TN+FP+FN)}
\]

TP is the number of samples which are truly detected positive similarly TN and FP is equal to the number of samples which are truly detected negative and the number of samples which are wrongly detected positive respectively. FN is to the number of samples which are wrongly detected negative (False negatives).

Results:
The table shows the different values obtained after the calculation:

<table>
<thead>
<tr>
<th>MODELS</th>
<th>PRECISION</th>
<th>RECALL</th>
<th>FMEASURE</th>
<th>ACCURACY</th>
</tr>
</thead>
<tbody>
<tr>
<td>SVM</td>
<td>88.3</td>
<td>90</td>
<td>86.53</td>
<td>81.77</td>
</tr>
<tr>
<td>KNN</td>
<td>78.12</td>
<td>84.95</td>
<td>81.39</td>
<td>79.16</td>
</tr>
<tr>
<td>NAIVE BAYES</td>
<td>80.85</td>
<td>79.49</td>
<td>80.16</td>
<td>75.52</td>
</tr>
<tr>
<td>ID3</td>
<td>77.40</td>
<td>71.15</td>
<td>74.14</td>
<td>66.40</td>
</tr>
<tr>
<td>C4.5</td>
<td>80.51</td>
<td>83.50</td>
<td>82.64</td>
<td>78.12</td>
</tr>
<tr>
<td>CART</td>
<td>76.98</td>
<td>78.63</td>
<td>77.80</td>
<td>72.65</td>
</tr>
</tbody>
</table>

According to the above evaluation on PIMA dataset, SVM gave the best results as the % accuracy was 81.77. Therefore the SVM model was the most accurate model for diagnosing diabetes. In addition C4.5 decision tree is one of the most famous method and proved to be more accurate as compared to ID3 and CART.

2.2 Data Warehousing

Before being transferred to the Hadoop environment the massive and unstructured data is cleansed, gathered and made ready for further processing.

2.3 Hadoop Environment

Patient’s needs can be accurately responded by Predictive Analysis in health care sector by easy financial and clinical decision making ability based on the predictions made by the system. Here Hadoop-Map Reduce environment is used to predict the complications associated to Diabetes Mellitus and the treatment to be given. Hadoop, is an open source distributed data processing platform. It has the potential to process extremely large amounts of data. Hadoop uses two main components to do its job:

- Map Reduce: Hadoop’s Map Reduce technique processes data sets by dividing them into small blocks of tasks. It uses distributed algorithms on a group of computers in clusters to process large data sets. The building block of Map Reduce algorithm is its two functions:
  - Map () – It divides the input data/task into smaller tasks at the Master node, then distributes the work to the Slave nodes, which evaluates smaller tasks and passes the results to the master node. The subtasks are performed parallel on various systems.
  - Reduce () – it collects the results of all the subtasks and combines them to produce an aggregate final result.
- HDFS: The data blocks that resides on various other computers is replicated by HDFS to ensure readability and manages the transfer of data from various parts of the distributed system.

2.4 Predictive Pattern Matching
Fig 2: Predictive pattern matching using map-reduce

Patterns of diabetic data includes plasma glucose concentration, serum insulin, diastolic blood pressure, diabetic pedigree, BMI etc. The predictive pattern matching is the process of comparison of analyzed edge point with the obtained value. After the data is warehoused it is sent to the Hadoop system, where map-reduce task is immediately performed. During the mapping phase, the Master node splits large data sets into smaller tasks for numerous worker nodes. The worker nodes receives the tasks from the master node, processes the pattern matching task for diabetic data with the help of Data Node of the same machine and the task tracker. If pattern matching is completed by all the node it is stored in intermediate disks. Based on the query from client through master node the reduce operation takes place on the worker node and finally the results obtained from reduce phase are distributed on various servers.

CHAPTER 3: PREDICTION OF DIABETIC RETINOPATHY USING SVM ALGORITHM

Diabetic retinopathy (DR) is testified as the most recurrent cause of blindness worldwide in individuals. During the first two decades of diabetes, nearly all patients with type1 diabetes and over 60% of patients with type2 diabetes have retinopathy. Its occurrence is directly proportional to the duration of Diabetes Mellitus. The occurrence of Diabetic retinopathy is initiated mild non-proliferative anomalies, which are distinguished by improved vascular permeability, vascular closure, severe non-proliferative diabetic retinopathy (NPDR) and finally ends with proliferative diabetic retinopathy (PDR), which is roots to growth of new blood vessels on the retina and the posterior surface of the vitreous. The high blood glucose level in blood is the cause of diabetic retinopathy. Therefore the blood sugar level should be maintained such a way that it does not end up to Diabetic retinopathy. Detection of such a disease at an early stage is difficult because of the huge amount of discrete data of each patient. The only solution for this is building of an involuntary recognition system.

Digital non-mydriatic images is preferred for automated system towards the investigation and discovery of diabetic retinopathy. This section gives a detailed review on the algorithm proposed to improve the performance and accuracy of classification.

3.1 METHODOLOGY

3.11. Retinal image procurement

To capture the retinal images Non-mydriatic retinal cameras are used in the medical. The detection of exudates is performed using an programmed screening system. The system requires non-mydriatic digital fundus photographs as the input. The MESSIDOR database consists of a collection of 1200 images, in which 800 images have a distention and 400 images do not have a distention. Using those 400 non-dilated images, selection of 370 digital fundus images was done in the age group of 25–65 years.

3.12. Preprocessing

In the preprocessing stage non-dilated low contrast retinal image is delivered as the input. The blurred image and the images with low clarity are rectified in the preprocessing step

(i) color space conversion
(ii) filtering
(iii) Contrast enhancement.

The translation of the represented color from one basis to another is called color space conversion. To carry out translation, the HSV (Hue (H), Saturation (S), and Value (V)) color space is used on the digital retinal fundus images of patients of MESSIDOR database. By examining the RGB color cube along the grey axis (the axis joining the black and white vertices) the HSV color space is generated, resulting in a hexagonally shaped color palette which is based on cylindrical coordinates, by this means making conversion from RGB to HSV similar to mapping the RGB coordinate values to the cylindrical coordinate’s function. Linear transition from high to low is made by hue values. The exudates can only be differentiated by intensity, therefore the intensity component alone is extracted from the HSV model.
Filtering- Before execution, a high degree of noise lessening in an image must be performed in higher level processing steps which is done by median filter, a nonlinear filtering technique that is used to eradicate clatter from images for improving the outcomes of later processing. The median filter is particularly useful to reduce the speckle noise and the salt and pepper noise. It is effective for suppressing isolated noise without blurring sharp edges since it is a simple edge-preserving smoothing filter.

Contrast enhancement To obtain a good quality of image the uneven illumination should be removed. To achieve this contrast enhancement, a method called histogram equalization is employed. A histogram delivers a appropriate summary of the strengths in an image and offers additional awareness concerning the image contrast and brightness. Histograms are calculated by softwares which is implemented on hardware tools making them popular for real-time image processing. The CLAHE (Contrast limited adaptive histogram equalization) algorithm splits an image into contextual regions and applies the histogram equalization to every region. It provides a uniform distribution of used grey values and thus permits the hidden topographies of the image to become more visible. The original and enhanced images are shown in Fig 4 below:

Contrast enhancement To obtain a good quality of image the uneven illumination should be removed. To achieve this contrast enhancement, a method called histogram equalization is employed. A histogram delivers a appropriate summary of the strengths in an image and offers additional awareness concerning the image contrast and brightness. Histograms are calculated by softwares which is implemented on hardware tools making them popular for real-time image processing. The CLAHE (Contrast limited adaptive histogram equalization) algorithm splits an image into contextual regions and applies the histogram equalization to every region. It provides a uniform distribution of used grey values and thus permits the hidden topographies of the image to become more visible. The original and enhanced images are shown in Fig 4 below:

Fig 4: (a) Original image (b) Enhanced image

3.13 ROI (Region of interest) Extraction:
The unwanted portion of the image is removed in the preprocessing step. Here in the segmentation process the ROI should be determined. The OD and the macula of the image has similar intensity as the exudates, therefore ROI containing macula and OD is eliminated.

3.14 OD (optical disc) elimination

The detection of optical disc (the starting point of optic nerve in eyes) is a primary step in the development of a computerized screening system for diabetic retinopathy. The OD appears as a bright yellowish or white region in the fundus images. The OD must be removed before the scrutinizing process because the detection of exudates is the main purpose of the screening. The largest high contrast areas is marked by OD. The blood vessels also appear with high contrast, but the sizes of these areas are much smaller than the area of the OD. Thus, the OD and macula are detected and masked. Mask generation is aimed at pointing the pixels belonging to the region of interest in the entire image. The masking process can be achieved by blurring the original image and subtracting the grubby image from the original image. Then, the mask is added to the original image. Thus an OD-eliminated image can be obtained.

3.21. Edge detection

The procedure through which sharp discontinuities are categorized and sharp discontinuities are retained in an image is called edge detection. The cutoffs are instant variations in the pixel concentration that distinguish the boundaries of objects. It is a fundamental tool for image segmentation. It transmutes the original images into edge images obtained from the changes of the grey tones in the image.

Kirsch edge detection removes blood vessels by using threshold values. The operator of kirsch is a nonlinear edge indicator that defines the maximum edge strength in a few pre-determined directions. The masks of this Kirsch technique considers a single mask and rotates it into eight main compass direction: north, northwest, west, southwest, south, southeast, east and northeast. The presence of edge and the direction of edge can be detected Kirsch edge detection algorithm. A convolution table stores a pixel \( [i,j] \) in Kirsch edge detection using 3*3 table.

3.22 Image segmentation

The method used for the detection of exudates that is performed by connected component labelling based on a neighborhood approach is score computation. The aim of the neighborhood connected component labelling analysis is to identify the large-sized bonded foreground region
in an image. The relevant region is identified using the neighborhood component labelling operation. The labelling can be performed by assigning a single label to each maximally connected foreground region of pixels. It is very useful to remove the regions that are not separated by a boundary, once the region boundaries are detected. If any pixel set is not separated by a boundary, then it is referred to as cell connected. Each maximal region of the connected pixels is called a connected component. A set of connected components partitions an image into segments. Eight neighbor connected component analysis is used for detecting exudate pixels by discarding all other pixels. This analysis connects the neighbor pixel to the center pixel which has the same intensity value by labelling them. The connected components labelling operator scans the image by moving end to end a row until it comes to point p (where p denotes the pixel to be labelled at any stage in the scanning process), for which V = {1}. When it reaches point p the operator scrutinizes the four neighbors of p, which have already been crossed in the scan (i.e., the neighbors (i) to the left of p, (ii) above p, (iii) one of the upper diagonal terms, and (iv) the next upper diagonal term). The labelling of p occurs as follows:

I. If all four neighbors are 0 → a new label is assigned to p
II. If one neighbor has V={1} → assign its label to p
III. If more than one neighbor has V={1} → assign one of the labels to p and note the equivalences.

A unique label is assigned to each class. A class is formed by sorting equivalent label pairs. Each label is replaced by the label assigned to its equivalence classes in the final scan of the image.

Labelling of neighbor connected component is initiated by determining the non-background neighbors. If any of the neighboring pixels is not labelled yet, then the label count is increased and the pixel is set to the current pixel; also, it is set as the parent to itself. Next the algorithm moves to the next pixel that has a neighbor, which is already labelled, and assigns the pixel’s label to that of the neighbor. All the neighboring pixels are labelled. The label count is thus increased and assigned to the pixel and again located its parent in relation to itself. When neighbors have dissimilar labels, any one of the labels is selected and is set to the current pixel. In this way moderately and severely affected samples.

### 3.3 Feature extraction

The patterns observed in the image provides information about the image, these patterns are called features. The feature extraction stage determines the accuracy of the classification. These features are intelligent enough to isolate normal and abnormal scratches with masses and microcalcification. The Grey Level Co-occurrence Matrix (GLCM) extracts the input required for SVM classifier. The matrix contains information about the positions of pixels having similar grey level values. The matrix can make use of the distance vector. The GLCM is represented as G[i, j]. The distance vector having grey levels at i and j is used to calculate all of the pairs of pixels. Based on the investigated matrix and the texture information, the parameters, such as contrast, correlation, cluster prominence, cluster shade, energy, homogeneity, entropy and maximum probability, are obtained. The extracted features are applied to the SVM classifier to determine the severity of the disease. All of the features are combined and given to the classifier to reduce the duplication, the improbability and the obscurity so that the classifier can provide grander classification.

![Fig 6: (a) moderately affected original image and (b) segmented image.](image)

![Fig 7: (a) Severely affected original image and (b) segmented image.](image)

### 3.4 SVM classification

The SVM classifier is used to separate normal and abnormal images, as a result, a good performance can be achieved because SVM minimizes the
empirical risk and prevents the over-fitting problem. After segmenting the regions of the exudates in the color retinal images, the segmented image is processed using the SVM classifier. This classifier is used to evaluate the training data to determine the best approach to classify images into different cases, such as moderate or severe.

Based on the reviewing paper statistical analysis of the extracted features is represented in the following table.

<table>
<thead>
<tr>
<th>Features of images</th>
<th>Normal</th>
<th>Abnormal (Moderate)</th>
<th>Abnormal</th>
</tr>
</thead>
<tbody>
<tr>
<td>Contrast</td>
<td>0-0.1</td>
<td>0.1-0.15</td>
<td>0.15-0.2</td>
</tr>
<tr>
<td>Correlation</td>
<td>0.9-0.94</td>
<td>0.95-0.97</td>
<td>0.97-0.99</td>
</tr>
<tr>
<td>Cluster prominence</td>
<td>&lt;100</td>
<td>101-200</td>
<td>&gt;200</td>
</tr>
<tr>
<td>Cluster stable</td>
<td>&lt;=0</td>
<td>0.1-5</td>
<td>5.1-10</td>
</tr>
<tr>
<td>Energy</td>
<td>0.2-0.3</td>
<td>0.30-0.35</td>
<td>0.35-0.40</td>
</tr>
<tr>
<td>Entropy</td>
<td>8.1-8.7</td>
<td>8.8-9.3</td>
<td>9.4-9.9</td>
</tr>
<tr>
<td>Homogeneity</td>
<td>0-10</td>
<td>10.01-15</td>
<td>&gt;15.01</td>
</tr>
<tr>
<td>Max. Probability</td>
<td>0-0.1</td>
<td>0.10-0.15</td>
<td>0.15-0.2</td>
</tr>
</tbody>
</table>

This classifier evaluates the training data to determine the best approach to classify images into different cases, such as moderate or severe. The classification was performed using MATLAB software, where the conditional programming method is used for classification. The relevant features of each and every normal and abnormal image are extracted. Next, the mined features of all of the regular and anomalous images are combined to form a matrix called a matrix file. This matrix file was the input to the classifier, and then, the training mode is also saved as a matrix file because it is required for the classifier testing mode. Successively, for testing an image, the relevant features of the test image are extracted and combined to form a matrix file. Before this matrix file is input into the SVM classifier, the matrix file of the training mode was loaded to determine the severity of the disease.

**CONCLUSION:**

Diabetes has become one of the most dangerous and most exorbitant diseases in the world that is currently rising at a frightening rate. Data mining algorithms are operative in designing appropriate models for diagnosing diabetes. This review paper concludes that SVM model was the most accurate of all and can be useful in inspection of diagnosing people inclined to danger. Big Data Analytics in Hadoop’s implementation provides orderly way for accomplishing better outcomes like obtainability and affordability of healthcare facility to all people. By transmuting various fitness records of diabetic patients to useful analyzed result, can make the patient understand the difficulties to occur. Prediction of Diabetic Retinopathy can be helpful in saving millions of sights to move towards the life of darkness and disparity.

**REFERENCES:**


2. Urmia University of Medical Science, Urmia, Iran. Department of Computer, Boukan Branch, Islamic Azad University, Boukan, Iran. 3 Young Researchers and Elite Club, Urmia Branch, Islamic Azad University, Urmia, Iran. Data Mining Algorithms Application in Diabetes.

3. Dr Saravana Kumar N M1 *, Esvari T2, Sampath P3 & Lavanya S4 Associate Professor, Dept of CSE, Bannari Amman Institute of Technology, Sathyamangalam – 638401, India. Predictive Methodology for Diabetic Data Analysis using Big Data.


