International Journal of Scientific Research in Computer Science, Engineering and Information Technology
ISSN : 2456-3307 (www.ijsrcseit.com)
doi : https://doi.org/10.32628/CSEIT217425

Analysis And Detection of Diabetes Using Data Mining Techniques – Efficiency Comparison

G. Ramadevi[1], Srujitha Yeruva[2], P. Sravanthi[2], P. Eknath Vamsi[2], S. Jaya Prakash[2]

[1]Assistant Professor, Department of Computer Science and Engineering, VVIT, Guntur, Andhra Pradesh, India

[2]Department of Computer Science and Engineering, VVIT, Guntur, Andhra Pradesh, India

Article Info
Volume 7, Issue 4
Page Number : 73-79

Publication Issue :
July-August-2021

Article History
Accepted : 02 July 2021
Published : 08 July 2021

ABSTRACT

In a digitized world, data is growing exponentially and it is difficult to analyze the data and give the results. Data mining techniques play an important role in healthcare sector - BigData. By making use of Data mining algorithms it is possible to analyze, detect and predict the presence of disease which helps doctors to detect the disease early and in decision making. The objective of data mining techniques used is to design an automated tool that notifies the patient's treatment history disease and medical data to doctors. Data mining techniques are very much useful in analyzing medical data to achieve meaningful and practical patterns. This project works on diabetes medical data, classification and clustering algorithms like (OPTICS, NAIVEBAYES, and BRICH) are implemented and the efficiency of the same is examined.

Keywords : Big data health care, Data mining techniques, Gaussian Naïve Bayes, OPTICS, BIRCH

I. INTRODUCTION

In India, healthcare systems have gained importance in recent years with the emergence of Big Data analytics Diabetes mellitus is posing a unique health problem in the country today, and hence India ranks top in the world. Diabetes is a chronic medical condition that can be administered and controlled through changes in lifestyle at an initial stage. At advanced state, diabetes can be controlled easily with early time detection and proper medication. Statistics as of today quotes that approximately 145 million people worldwide are affected by diabetes mellitus and 5% of the Indian population contributes towards this rate. Diabetes is a condition during which the physical body won't be ready to generate the available amount of insulin which is important to balance and monitor the amount of sugar in the body. Severe stage of diabetes can also lead to heart diseases, blindness, kidney failure etc. Diabetes depends on two reasons:

- Required amount of Insulin is not produced by the pancreas. This specifies Type-1 diabetes and occurs in 5–10% of people.
- In Type-2, insulin production cells become inactive. Gestational diabetes usually attacked in women when a high sugar level is

generated during pregnancy.

Table 1 : Comparison between type1 and type2 diabetes

Feature	**Type 1 diabetes**	**Type 2 diabetes**
Onset	Sudden	Gradual
Age	children	adults
Body size	Thin or normal	Often obese
Ketoacidosis	Common	Rare
Autoantibodies	Usually present	Absent
Prevalence	~10%	~90%

Interpretation and analyzing the presence of diabetes is a significant problem to classify. The Classifier is intended such that it is more convenient and cost-efficient. Big Data and data mining techniques provide a great deal to human-related applications. These methods find the most appropriate space in the medical diagnosis which is one of the classification phenomena. A physician is supposed to analyze many factors before the actual diagnosis of the diabetes leading to a difficult task. Designing of automated diabetic detection uses machine learning and data mining techniques.

II. MACHINE LEARNING

Machine learning (ML) is the study of computer algorithms that automatically improves the efficiency of complex tasks. It is seen as a subset of artificial intelligence. Machine learning algorithms build a mathematical model based on sample data, known as "training data", in order to make predictions or decisions without being explicitly programmed to do so. Machine learning algorithms applications are used in filtering emails, computer vision and it is difficult to develop complex algorithms to perform required tasks.

A. Relation to Data Mining:

Machine Learning and Data Mining approach the same methods, but Machine Learning focuses on prediction learned from the Training data. Data Mining focuses on the unrecognized properties in the data. Data Mining combines with Machine Learning methods, but goals vary. Machine Learning also uses Data mining methods such as 'Unsupervised Learning' to improve learning methodologies.

B. Relation to optimization:

Machine Learning also deals with optimization. Loss function on a training set is a set of examples. Loss functions describes the inconsistency between the predictions of the model being trained and the actual problem instances. The difference between the fields are from the goal of generalization optimization algorithms can minimize the loss on a training set whereas machine learning is concerned with minimizing.

C. Relation to statistics:

Machine learning and statistics were closely related fields in terms of, but distinct in their principal goal: statistics gets population inferences from a sample, where as machine learning identifies generalizable predictive patterns. According to Michael I. Jordan, the ideas of machine learning, from methodological principles to theoretical tools, have had an extended pre-history in statistics. The term data science as a placeholder to call the overall field.

III. METHODOLOGY

Data mining is a new pattern for analyzing medical data and achieving useful and practical patterns. Data mining helps us to predict the type of disease and tries to find already non-identified patterns. The objective of the proposed methodology is to analyze the medical dataset and predict whether the patient

is suffering from diabetes disease or not. The prediction for diabetes is done using data mining algorithms such as Gaussian Naïve Bayes, BIRCH and OPTICS. The Naïve Bayes technique is applied to the dataset to expect whether the patient is diabetic or non-diabetic. BIRCH and OPTICS clustering algorithms are used to cluster people with similar disease into one cluster and identify which algorithm is more efficient by calculating the efficiency measures.

A. Input Dataset

The Dataset used for the application is the "Pima Indian diabetes dataset". The dataset consists of several medication predictor(independent) variable and one target(dependent) variables, outcome. The dataset is a CSV(Comma Separated Value) file. It contains upto 760 records. This dataset is taken from National Institute of Diabetes and Digestive and Kidney Diseases. The main objective of the dataset is to predict whether the patient is having diabetes or not, based on available diagnostic measurements included in the dataset. Several conditions were placed for the selection of these instances from a huge database. In this dataset all patients were females of age 21 years old of Pima Indian heritage.

Features in dataset are:

- Number of times pregnant
- Plasma glucose concentration
- Diastolic blood pressure
- Triceps skin fold thickness
- 2-Hour serum insulin
- Body mass index
- Diabetes pedigree function
- Age
- Outcome

For the dataset, we apply algorithms to detect whether a patient is diabetic or not. The dataset consists of nine features with a class variable called the outcome variable.

IV. ALGORITHMS USED IN PROPOSED SOLUTION

A. Gaussian Naïve Bayes

Naïve Bayes classifiers are simple probabilistic classifiers based on seeking Bayes' theorem with strong independence and assumptions between the features.

Why do we use Naïve Baye's :- simple and easy to implement , Doesn't require training data, Highly Scalable, fast and can be used to make real-time predictions, It is not sensitive to irrelevant features.

Naïve Bayes classifier works on the principle of conditional probability given by Bayes Theorem. Bayes' theorem allows updating the estimated probabilities of an event by including new information.

B. Optics Algorithm

OPTICS Algorithm is abbreviated as Ordering Points to Identify Cluster Structure. It updates from the DBSCAN clustering algorithm. Two more terms are updated to optics from DBSCAN clustering. They are

1) Core Distance:

Core Distance is the minimum value of radius which are essential to classify a given point as a core point. If the given point is not a Core point, then its Core Distance is undefined.

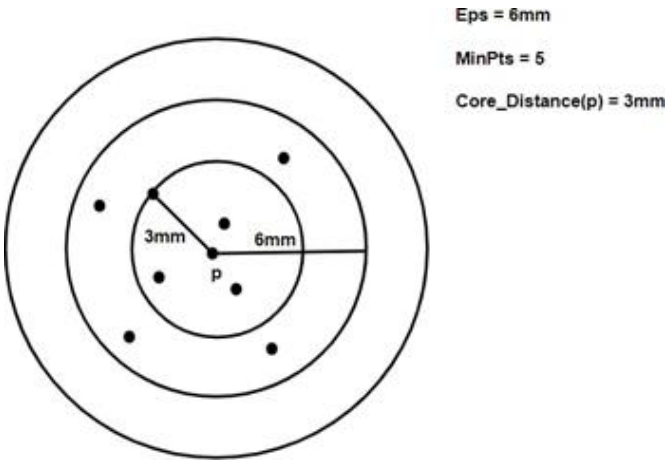

Fig. 1: Core distance

2) Reachability Distance:

This clustering technique is different from other techniques such that this technique does not explicitly branch the data into clusters. Visualization of Reachability distances is produced and is used to cluster the data.

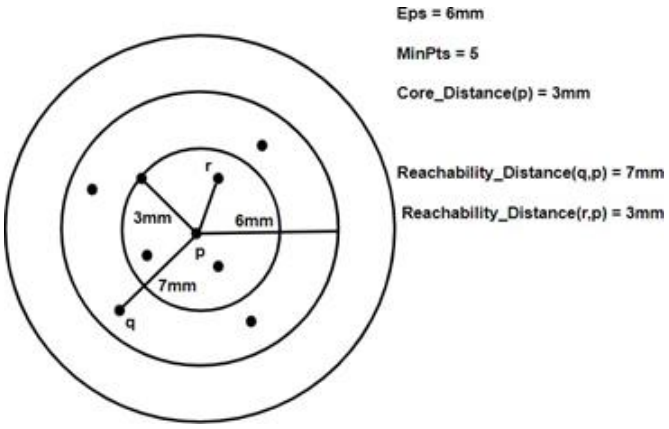

Fig. 2: Reachability distance

3) Algorithm steps:

Step 1: Initially ε and MinPts got to be specified.
Step 2: All the data points with data in the dataset are marked as unprocessed.
Step 3: Neighbors are found for each point p which is unprocessed.
Step 4: Now mark the data point as processed.
Step 5: Configure the core distance to the data point p.
Step 6: Create an Order file and include data point p in the file.
Step 7: If core distance initialization is unsuccessful, return back to Step 3 otherwise visit
Step 8: Calculate the reachability distance for each of the neighbors and update the order seed with the reference of the latest values.
Step 9: Find the neighbors for each data point in order and update the point as processed.
Step 10: Fix the core distance of the point and append the order file.
Step 11: If there is an undefined core distance, go to Step 9, else continue with Step 12.
Step 12: Repeat Step 8 until no change in the order
Step 13: End.

C. Birch Algorithm

Balanced Iterative Reducing and Clustering using Hierarchies (BIRCH) is a clustering algorithm that is able to cluster large datasets by generating a small and compact summary of the large dataset which holds as much information as possible. The smaller summary is clustered instead of clustering the larger dataset. BIRCH is often used to complement other clustering algorithms by establishing a summing-up of the dataset that the other clustering algorithm can now use. BIRCH has one major drawback, only the metric attributes can be processed. A metric attribute is an attribute whose values are often represented in Euclidean space

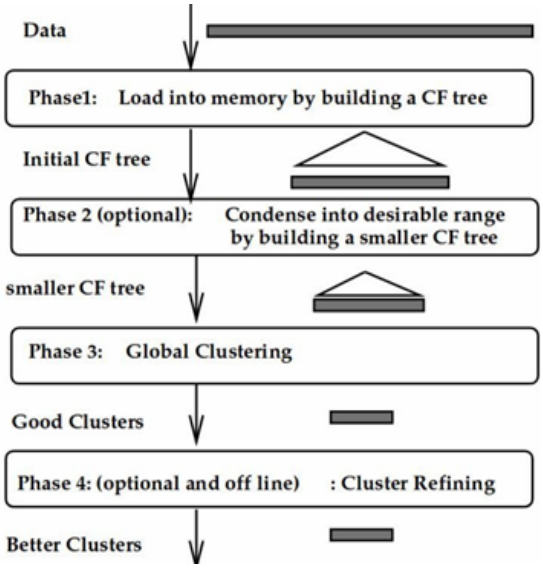

Fig. 3: Phases of BIRCH Algorithm

Phase 1: Scan the dataset and construct an initial in-memory CF tree.

Phase 2: Scan all the leaf entities of the CF tree and build a replacement CF tree that is smaller in size. Terminate all the outliers and form the clusters.

Phase 3: Use the clustering algorithm to cluster all the leaf entities. This phase leads to create a group of clusters.

Phase 4: The cluster centroids obtained in Phase 3 are used as seeds and the data points are redistributed to their closest neighbour seeds to form new cluster representations. Finally, each leaf entity signifies each cluster class.

1) Algorithm steps:

Step 1: Set an initial threshold value and insert data points to the CF tree with respect to the Insertion algorithm.

Step 2: Increase the edge value if the dimensions of the tree exceed the memory limit assigned to it.

Step 3: Reconstruct the partially built tree consistent with the newly set threshold values and memory limit. Step 4: Repeat the above steps until all the data objects are scanned which forms a complete tree.

Step 5: Smaller CF trees are built by varying the edge values and eliminating the Outliers.

Step 6: Considering the leaf entities of the CF tree, the cluster quality is improved accordingly by applying the universal clustering algorithm.

Step 7: Redistribution of data objects and labelling each point in the completely built CF tree.

V. COMPARISON BETWEEN PERFORMANCE OF ALGORITHMS

The performance of algorithms is calculated by using precision, recall and F1 scores.

Precision: Precision is a good measure to determine when the values of False Positive is high. For instance, email spam detection. In email spam detection, a false positive means an email which is non-spam (actual negative) has been identified as spam (predicted spam). The user might lose important data if the precision is not high for the spam detection model.

Recall: Recall actually calculates what percent of the Actual Positives our model capture through labelling it as Positive (True Positive). Applying the equivalent understanding, we know that Recall shall be the model metric we use to select our greatest model when there is a high cost related to False Negative.

F1 Score: F1 Score is needed when you want to seek a balance between Precision and Recall. what will be the difference between the F1 Score and Accuracy then? We have previously seen that accuracy is often largely contributed by a huge number of True Negatives which is mostly observed in business circumstances, we do not focus on much whereas False Negative and False Positive usually consists of business costs.

Table 2 : Comparison between Optics and BIRCH

ALGORITHM	PRECISION	RECALL	F1 SCORE
Optics	0.59	0.59	0.59
BIRCH	0.42	0.415	0.40

Based on the above performance metrics table of the two clustering algorithms Optics and BIRCH, the best algorithm that is most suitable for Diabetes detection is the Optics algorithm. Here a comparison is considered between the algorithms that are specified above and in terms of all the parameters considered Optics is considered as the best algorithm.

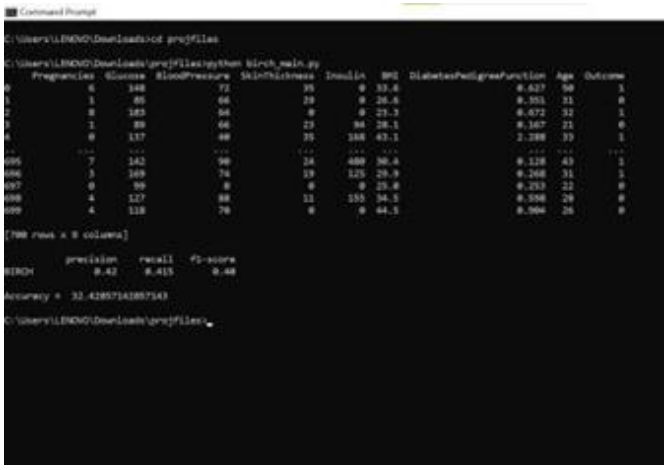

Fig. 4: Accuracy of BIRCH

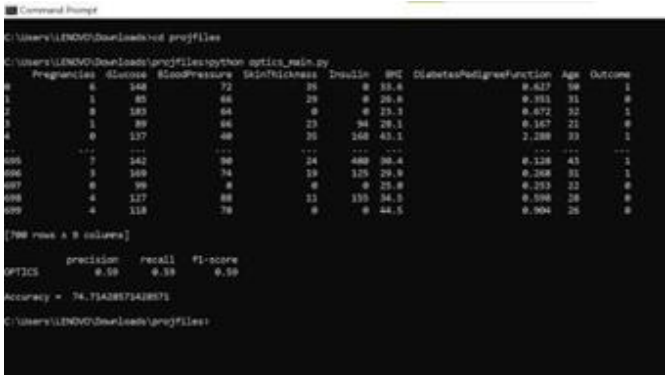

Fig. 5: Accuracy of OPTICS

VI. CONCLUSION AND FUTURE SCOPE

The usefulness of data mining algorithms like Gaussian Naïve Bayes, BIRCH and OPTICS for the prediction of diabetic disease is demonstrated. Data mining techniques are constructive in diagnosing and clustering the report of diabetic patients. BIRCH and OPTICS are used to cluster similar kinds of people, where BIRCH deploy on the CF tree and OPTICS deploy on the ordering of the points in the cluster. Analysis and comparison of clustering algorithms are executed by considering numerous performance metrics. It is observed that for the same number of clusters obtained by different clustering techniques, OPTICS is the most efficient and is suitable for diagnosis of diabetes. This work helps the doctors to diagnose and supply the recommended medicine at an early stage to the patient to cure the disease. The main aim is to reduce the cost and provide better treatment. In future, this can be worked with an additional number of classification algorithms and their accuracy can be compared to find the optimal one.

VII. REFERENCES

[1]. Raghupathi, W., & Raghupathi, V. (2014). Big data analytics in healthcare: Promise and potential. Health Information Science and Systems.

[2]. Diabetes Mellitus https://en.wikipedia.org/wiki/Diabetes_mellitus

[3]. Agicha, K., et al. Survey on predictive analysis of diabetes in young and old patients. International Journal of Advanced Research in Computer Science and Software Engineering.

[4]. Iyer, A., Jeyalatha, S., & Sumbaly, R. (2015, January). Diagnosis of diabetes using classification mining techniques. International Journal of Data Mining & Knowledge Management Process (IJDKP), 5(1).

[5]. Ankerst, M., Breunig, M. M., Kriegel, H.-P., & Sander, J. OPTICS: Ordering Points To Identify the Clustering Structure. Institute for Computer Science, University of Munic.

[6]. Alzaalan, M. E., & Aldahdooh, R. T. (2012, February). EOPTICS "Enhancement ordering points to identify the clustering structure". International Journal of Computer Applications (0975–8887), 40(17).

[7]. Senthil kumaran, M., & Rangarajan, R. (2011). Ordering points to identify the clustering structure (OPTICS) with ant colony optimization for wireless sensor networks. European Journal of Scientific Research, 59(4), 571–582 (ISSN 1450-216X).

[8]. Zhang, T., Ramakrishnan, R., & Livny, M. (1997). BIRCH: A new data clustering algorithm and its applications. Data Mining and Knowledge Discovery, 1, 141–182.

[9]. Zhang, T., Ramakrishnan, R., & Livny, M. BIRCH: An efficient data clustering method for very large databases.

[10]. Du, H. Z., & Li, Y. B. (2010). An improved BIRCH clustering algorithm and application in thermal power. In 2010 International Conference on Web Information Systems and Mining.

[11]. Feng, X., & Pan, Q. The algorithm of deviation measure for cluster models based on the FOCUS framework and BIRCH. In Second International Symposium on Intelligent Information Technology Application.

[12]. UCI Machine Learning Repository Pima Indians Diabetes Database https://archive.ics.uci.edu/ml/datasets/Pima+India ns+Diabetes.

[13]. Naive Bayes. https://en.wikipedia.org/wiki/Naive_Bayes_classif ier.

[14]. Optics Algorithm. https://en.wikipedia.org/wiki/OPTICS_algorithm..

[15]. Birch Algorithm. https://people.eecs.berkeley.edu/~fox/summaries/d atabase/birch.html

Cite this article as :

G. Ramadevi, Srujitha Yeruva, P. Sravanthi, P. Eknath Vamsi, S. Jaya Prakash, "Analysis And Detection of Diabetes Using Data Mining Techniques – Efficiency Comparison", International Journal of Scientific Research in Computer Science, Engineering and Information Technology (IJSRCSEIT), ISSN : 2456-3307, Volume 7, Issue 4, pp.73-79, July-August-2021. Available at
doi : https://doi.org/10.32628/CSEIT217425
Journal URL : https://ijsrcseit.com/CSEIT217425