Parameter Tuning of DBSCAN for Medical Data and Diabetes Diagnosis

[Extended Abstract]

Michael R. Wyatt II
University of Delaware
18 Amstel Ave
Newark, DE 19701
mwyatt@udel.edu

Michela Taufer
University of Delaware
18 Amstel Ave
Newark, DE 19701
taufer@udel.edu

ABSTRACT

The increasing use of computationally assisted diagnosis in the doctor's office requires that computer diagnosis be both fast and accurate. We present a scalable method for preparing laboratory data for use with learning algorithms and a method for identifying optimal parameter settings for learning algorithms. To demonstrate our method, we predict the presence of diabetes among participants of the National Health and Nutrition Examination Survey using collected laboratory data and the DBSCAN algorithm. We performed optimization of the DBSCAN parameters for this dataset to demonstrate how diagnosis predictions can be improved.

CCS Concepts

•**Applied computing** → **Health care information systems;** *Consumer health;* •**Computing methodologies** → *MapReduce algorithms;*

Keywords

Health Informatics; Machine Learning; Optimization

1. MOTIVATION

Modern medical diagnosis is becoming increasingly computationally assisted. This means that artificial intelligence and machine learning algorithms are being used to analyze patient medical data and provide a diagnosis. Human doctors consult the computation results and a final diagnosis is made. As computationally assisted diagnosis becomes more widespread, patient diagnosis becomes more accurate [1] [3]. An important aspect of computationally assisted diagnosis is the processing of medical data by learning algorithms to produce useful results. Improving the speed and accuracy of this process will encourage the continued adoption of computationally assisted diagnosis by medical doctors, which will lead to improved population health and disease management.

2. CONTRIBUTIONS

Many efforts have been made to improve both the accuracy and processing time of computationally assisted diagnosis. These efforts focus mainly on the application of different algorithms to medical data. In this paper, we apply the clustering algorithm DBSCAN to medical data in order to diagnose patients with diabetes. We present a framework for parallel processing of medical data for learning algorithms. Additionally, we outline a method for optimizing learning algorithm parameters to achieve optimal results.

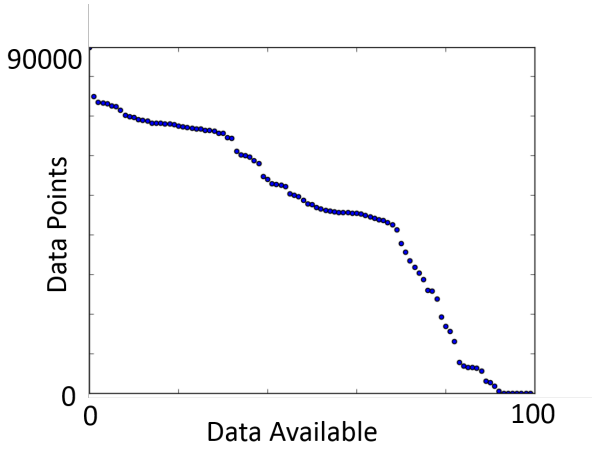

Figure 1: NHANES participants sorted by percent of laboratory data available.

3. METHODOLOGY

We isolate lab data from the National Health and Nutrition Examination Survey (NHANES) dataset for over 70,000 participants. We process this data using parallel algorithms built with the MapReduce programming paradigm via Apache Spark. The processing of data is highly scalable across many nodes. The processed data is in a form that is usable by learning algorithms like DBSCAN. We then perform parameter optimization to achieve maximum predictive capabilities.

3.1 NHANES Dataset

We obtained data from the NHANES continuous dataset collected between 1999 and 2014. We label the participants as having or not having diabetes based on their categorical response to the question, "have you ever been told by a doctor or health professional that you have diabetes or sugar diabetes?" We isolate 116 common features within the laboratory data, including urine and blood sample values, which can be used to cluster the diabetic and non-diabetic NHANES participants.

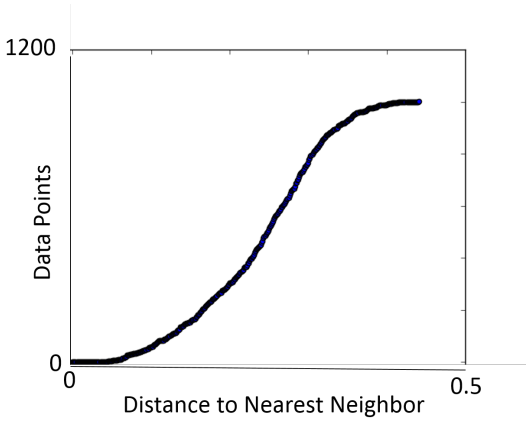

Figure 2: Number of data points (y axis) with a neighbor within distance *Epsilon* (x axis). The elbow in the plot indicates optimal values for *Epsilon*.

3.2 Data Preparation

We prepare the data by removing participants with many missing values, standardizing the data by feature, and dimensionality reduction. Figure 1 plots the participants sorted by the percent of data points (from the 116 features) available. All NHANES participants had missing values and many had most laboratory values missing. We identify a subset of 16,627 participants with more than 80% of features available for analysis. Each of the 116 features are standardized using Z-score standardization. This process makes the range of values for each feature similar to prevent one feature outweighing others (due to a large range of values). Dimensionality reduction is performed with Principal Component Analysis (PCA). The processing of NHANES data is performed with MapReduce algorithms. This allows the process to be distributed across many compute nodes and reduces result turn-around time.

3.3 Choosing Initial Parameters

There are three parameters for DBSCAN. Together, they define the density of clusters which will be found.

1. *Distance* - Metric used for calculating distance between patients
2. *Epsilon* - Distance around patients to identify neighboring patients
3. *Min_pts* - Number of neighboring patients to be "core" point

Like other learning algorithms, these parameters affect the quality of results. We chose to test two distance metrics: Euclidean and Cosine. We define cosine distance in equation 1. A range of *Min_pts* values was selected for testing. We then determined values for *Epsilon* by adapting the "elbow method" used for determining the best value of k in k-Means clustering [2]. Figure 2 shows the sorted Cosine distance to the nearest neighbor for each participant. We propose that ideal values for *Epsilon* will be around the elbow of this figure. In the case of figure 2, this range is $[0.15, 0.35]$.

$$Cosine\_distance(x, y) = 1 - \frac{x \cdot y}{\|x\|\|y\|} \quad (1)$$

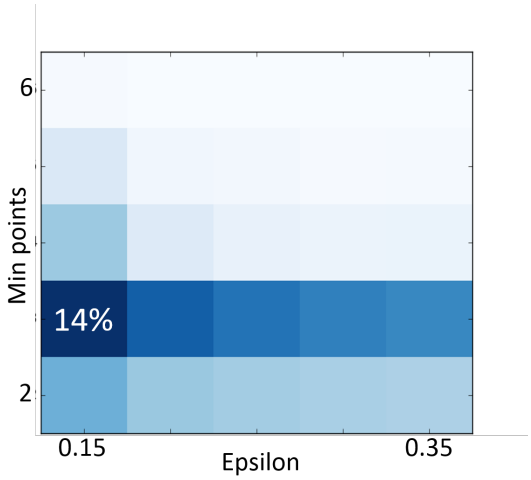

Figure 3: Information gain scoring method for Cosine distance metric across several *Epsilon* and *Min_pts* values.

3.4 Evaluation

To evaluate DBSCAN performance, we developed a scoring metric based on information gain. This scoring metric, seen in equation 2, considers the amount of information gain and size of each cluster in order to produce a value between 0 and 1. We scored each set of DBSCAN parameters and compared scores to find the best settings.

$$score = \frac{ClusteredPatients}{TotalPatients} \cdot \frac{InformationGain}{MaxInformationGain} \quad (2)$$

4. RESULTS

We observed that Cosine distance scored much higher than Euclidean distance. Figure 3 shows the scores of our parameter sweep around initial parameter selections for the Cosine distance metric. There is a clear set of parameters at $Epsilon = 0.15$ and $Min\_pts = 3$ which provides the best clustering score. The maximum observed score from our testing was 14%. Observed scores from our clustering were unexpectedely low. Despite the poor ability of DBSCAN to differentiate between patients with and without diabetes, our method of focused parameter searching was able to find optimal parameter settings.

5. REFERENCES

[1] K. Doi. Computer-aided diagnosis in medical imaging: historical review, current status and future potential. *Computerized medical imaging and graphics*, 31(4):198–211, 2007.

[2] T. M. Kodinariya and P. R. Makwana. Review on determining number of cluster in k-means clustering. *International Journal*, 1(6):90–95, 2013.

[3] I. Kononenko. Machine learning for medical diagnosis: history, state of the art and perspective. *Artificial Intelligence in medicine*, 23(1):89–109, 2001.