International Journal of Trend in Scientific Research and Development (IJTSRD)
International Open Access Journal
ISSN No: 2456 - 6470 | www.ijtsrd.com | Volume - 2 | Issue – 1

K-Nearest Neighbours based diagnosis of hyperglycemia

Abid Sarwar
Department of Computer Science & IT (Bhaderwah Campus)
University of Jammu, Jammu, India

ABSTRACT

AI or artificial intelligence is the simulation of human intelligence processes by machines, especially computer systems. These processes include learning (the acquisition of information and rules for using the information), reasoning (using the rules to reach approximate or definite conclusions), and self-correction. As a result, Artificial Intelligence is gaining Importance in science and engineering fields. The use of Artificial Intelligence in medical diagnosis too is becoming increasingly common and has been used widely in the diagnosis of cancers, tumors, hepatitis, lung diseases, etc... The main aim of this paper is to build an Artificial Intelligent System that after analysis of certain parameters can predict that whether a person is diabetic or not. Diabetes is the name used to describe a metabolic condition of having higher than normal blood sugar levels. Diabetes is becoming increasingly more common throughout the world, due to increased obesity - which can lead to metabolic syndrome or pre-diabetes leading to higher incidences of type 2 diabetes. Authors have identified 10 parameters that play an important role in diabetes and prepared a rich database of training data which served as the backbone of the prediction algorithm. Keeping in view this training data authors developed a system that uses the artificial neural networks algorithm to serve the purpose. These are capable of predicting new observations (on specific variables) from previous observations (on the same or other variables) after executing a process of so-called learning from existing training data (Haykin 1998).The results indicate that the performance of KNN method when compared with the medical diagnosis system was found to be 91%. This system can be used to assist medical programs especially in geographically remote areas where expert human diagnosis not possible with an advantage of minimal expenses and faster results.

Keywords: *Artificial Intelligence, metabolic, Machine Learning, Diabetes, K-nearest neighbors, Medical Diagnosis*

I. INTRODUCTION

Diabetes, often referred to by doctors as diabetes mellitus, describes a group of metabolic diseases in which the person has high blood glucose (blood sugar), either because insulin production is inadequate, or because the body's cells do not respond properly to insulin, or both. Patients with high blood sugar will typically experience polyuria (frequent urination), they will become increasingly thirsty (polydipsia) and hungry (polyphagia). Diabetes can mainly be of 3 types: Type-1 diabetes, Type-2 diabetes and Gestational diabetes. Type-1 diabetes results from non-production of insulin & Type-2 diabetes results from development of resistance of insulin, as a result of which the insulin produced is not able to metabolize the sugar levels properly. Gestational diabetes occurs in pregnant women, who develop a high blood glucose level during pregnancy who never had any previous such history. It may be preceded by development of type-2 diabetes. As reported by WHO & International Diabetic Federation in year 2010, the toll of diabetic patients was 285 million and this number is expected to grow to 483 million by 2030. WHO estimates that between 2010 to 2030 there will be an increase of 69% in adult diabetic population in developing countries and 20%

increase of the same in developed countries. India having 50.8 million patients of this disease leads the world & is followed by China (43.2), United States (26.8). In 2010 diabetes caused 3.9 million deaths worldwide. The primary concern of AI in medicine is the construction of AI programs that can assert a medical doctor in performing expert diagnosis. These programs by making use of various computational sciences such as statistics and probability find out the hidden patterns from the training data and using these patterns they classify the test data into one the possible categories. The backbones of these AI programs are the various data sets prepared from various clinical cases which act as practical examples in training the system. The decision and recommendation prepared from these systems can be illustrated to the subjects after combining them with the experience of human expert.

2. METHODOLOGY

K-nearest neighbor (KNN) is an instance-based classification algorithm which classifies the test objects on the basis of number of closest training examples. It is a nonparametric algorithm which means that it does not make any assumption on the underlying data distribution. It is also classified as a lazy learning algorithm. In this method, the Euclidean distance is calculated between the test data and all the samples in the training data which is followed by classifying the test data into a class which most of k-closest neighbors of training data belong to. K is a positive integer usually very small. Larger the value of K more difficult, it is to distinguish between the various classes. Various heuristic techniques such as cross-validation are used to make a good choice of K. This algorithm assumes that all the data correspond to points in N-dimensional space. Letthe test datumxi be represented by a feature vector [xi1, xi2,xi3….xin], where xik represents the kth attribute of test datum xi and xi0 represents the transpose of xi. The distance between xi and xj is defined by

$$d(xi, xj) = \sqrt{\sum (xj - xi)^2}$$

If the number of training samples is equal to n, then n such distances will be calculated and the closest K training data are identified as neighbors. If K = 1, then the class label of test datum is equal to the closest training datum. If there is a tie, then the tie is resolved arbitrarily

3.PREVIOUS WORK

It has been noted that the machine learning algorithms are increasingly being used in solving problems in Medical Domains such as in Oncology (Lisboa and Taktak 2006; Catto et al. 2003; Anagnostou et al. 2003; Bratko and Kononenko 1987), Urology (Anagnostou et al. 2003), Liver Pathology (Lesmo et al. 1983), Cardiology (Catlett 1991; Clark and Boswell 1991), Gynecology (Nunez 1990),Thyroid disorders (Hojker et al. 1998; Horn et al. 1985), and Perinatology (Kern et al. 1990). Numerous researchers have used AI and data mining-based algorithms to solve problems in the field of medicine. Bharat Rao et al. (2007) have proposed an artificially intelligent system (LungCAD) thathelps in the detection of lung cancer. They have applied a classification algorithm for detecting solid pulmonary nodules from CT thorax studies. The LungCAD system was clinically tested by a number of radiologists and was found to deliver significantly greater accuracy both in detecting the affected nodules and in identifying the potentially actionable nodules. LungCAD was approved by FDA in 2006.

4.PARAMETERS USED IN ESTIMATION

Since India is having the highest Diabetic population in the world so it was easy to collect the data about the patients who suffered from this disease. After a detailed study, authors identified ten best physiological parameters for the study which were so chosen that the values for them could be easily determined and could be assigned discrete values, for the sake of maintaining consistency. Table-I summaries the parameters chosen and their allowed values. A dataset of 415 cases was prepared by collecting the data randomly from different sections of the society with an aim to have a variety in the dataset. To maintain accuracy and to avoid errors, considerable care was taken to ensure that the database had correct values.

Table-I: Various parameters used and their allowed values

Parameter	Description	Allowed Values
Age	Age of the subject	Discrete Integer Values
Family History	Whether any family member of the subject is suffering/ was suffering from diabetes.	Yes or No
Sex	Whether male or female	Male or Female
Smoking	Whether the subject does smoking or not.	Yes or No
Drinking	Whether the subject does drinking or not.	Yes or No
Fatigue	Does a person feel tired after doing a little work?	Yes or No
Thurst	Whether the subject frequently feels a strong desire to drink water. i.e how many times the subject drinks water.	Discrete Integer values
Frequency of urination	How many times the subject passes urine in a day	Discrete Integer values
Height	Height of the subject	Discrete floating point values
Weight	Weight of the subject	Discrete floating point values

5. IMPLEMENTATION

For implementing the K-nearest neighbor algorithm, an application was developed in Matlab 7.6.0. The data collected for the study were divided into two parts viz training data and testing data which comprised of 90 and 10 %, respectively. The value of K was taken to be 5. Euclidean distance was calculated between the test data and each of the training data. The distance matrix thus obtained was then sorted. After sorting, we selected the first k items from the distance matrix, and these are those entries in trainingdata that are most likely to have the same symptoms of diabetes as that of our test data. This was followed by taking the votes of these k items, and the test data were classified into one of the corresponding category, that is, diabetic and not diabetic.

Fig. 1 MATLAB Program in execution

6. CONCLUSION

This K-nearest neighbor based system is very useful for diagnosis of diabetes. The reliability of the system was evaluated by predicting new observations (on specific variables) from previous observations (on the same or other variables) after executing a process of so-called learning from existing training data. The results suggest that this system can perform good prediction with least error and finally this technique could be an important tool for supplementing the medical doctors in performing expert diagnosis. In this method the efficiency of Forecasting was found to be around 91%. Its performance can be further improved by identifying & incorporating various other parameters and increasing the size of training data.

REFERENCE

1) Bharat Rao R, Bi J, Obuchowski N, Naidich D (2007) LungCAD: a clinically approved, machine learning system for lung cancer detection. In: International conference on knowledge discovery and data mining 2007, San Jose, California, USA, ACM 978-1-59593-609-7/07/0008

2) Bonham GS, Brock DB (1985) The relationship of diabetes with race, sex, and obesity. Am J ClinNutr 41:776–783

3) Bratko I, Kononenko I (1987) Learning rules from incomplete and noisy data. In: Phelip B (ed) Interactions in artificial intelligence and statistical methods. Technical Press, Hampshire

4) Caruana R, Niculescu-Mizil A (2006) An empirical comparison of supervised learning algorithms. In: ICML '06 Proceedings of the 23rd international conference on machine learning, pp 161–168

5) Catlett J (1991) On changing continuous attributes into ordered discrete attributes. In: Proceedings of European working session on learning-91, Portugal, 4–6 Mar 1991, pp 164–178

6) Catto JWF, Linkens DA, Abbod MF, Chen M, Burton JL, Feeley KM, Hamdy FC (2003) Artificial intelligence in predicting bladder cancer outcome: a comparison of neuro-fuzzy modeling and artificial neural networks. Clin Cancer Res 9:4172

7) Chan JM, Rimm EB, Colditz GA, Stampfer MJ, Willett WC (1994) Obesity, fat distribution, and weight gain as risk factors for clinical diabetes in men. Diabetes Care 17(9):961–969

8) Chang C-L,HsuM-Y(2009) The study that applies artificial intelligence and logistic regression for assistance in differential diagnostic of pancreatic cancer. Expert SystAppl 36(7):10663–10672

9) Chiu J-S, Wang Y-F, Su Y-C,Wei L-H, Liao J-G, LiY-C (2008)Artificial neural network to predict skeletal metastasis in patients with prostate cancer. Springer science ?Business Media, May 2008

10) Clark P, Boswell R (1991) Rule Induction with CN2: some recent improvements. In: Proceedings of European working session on learning-91, Portugal, Mar 1991, pp 151–163

11) Lesmo L, Saitta L, Torasso P (1983) Fuzzy production rules: a learning methodology. In: Advances in fuzzy sets, possibility theory and applications. pp 181–198

12) Lindley CA (2012) Neurobiological computation and synthetic intelligence. In: Artificial intelligence and simulation of behavior, AISB/IACAP world congress 2012, Birmingham, UK, 2–6 July 2012, pp 20–25

13) Mahamud K, Abu BakarA, Norwawi NM (1999) Multilayer perceptron modeling in housing market. Malays Manage J 3:61–69 Nunez M (1990) Decision tree induction using domain knowledge. In:

14) Wielinga B, Boose J, Gaines B, Schreiberg G, Someren Van M (eds) Current trends in knowledge acquisition. IOS Press, Amsterdam