International Journal of Advanced Computer Technology (IJACT)

ISSN:2319-7900

A Study on Mining Patient's Risk Factors in Diabetes using ID3 Algorithm

K.M. Padmapriya Asst. Professor, Dept. of Computer Science, SSM College of Arts and Science, Komarapalayam, India
N.Gowri Priya, M.Phil Research Scholar, SSM College of Arts and Science, Komarapalayam, India

Abstract

A Data mining technique is to analyze the data from different perspectives and summarizing it into useful information. This paper gives a Case Study, where the patient's Risk Factors is analyzed in concern with major causes of Diabetes using ID3 Algorithm. It helps earlier in identifying the people who need special attention in preventing them from diabetes with appropriate advising/counseling from the doctors.

Introduction

Diabetes is a set of related diseases in which the body cannot regulate the amount of sugar in the blood [1]. It is a group of metabolic diseases in which a person has high blood sugar, either because the body does not produce enough insulin, or because cells do not respond to the insulin that is produced. This high blood sugar produces the classical symptoms of polyuria, polydipsia and polyphagia [2]. There are three main types of diabetes mellitus (DM). Type 1 DM results from the body's failure to produce insulin, and presently requires the person to inject insulin or wear an insulin pump. This form was previously referred to as "insulin-dependent diabetes mellitus" (IDDM) or "juvenile diabetes". Type 2 DM results from insulin resistance, a condition in which cells fail to use insulin properly, sometimes combined with an absolute insulin deficiency. This form was previously referred to as non insulin-dependent diabetes mellitus (NIDDM) or "adult-onset diabetes". The third main form, gestational diabetes occurs when pregnant women without a previous diagnosis of diabetes develop a high blood glucose level. It may precede development of type 2 DM. As of 2000 it was estimated that 171 million people globally suffered from diabetes or 2.8% of the population. Type-2 diabetes is the most common type worldwide [3]. Figures for the year 2007 show that the 5 countries with the largest amount of people diagnosed with diabetes were India (40.9 million), China (38.9 million), US (19.2 million), Russia (9.6 million), and Germany (7.4 million) [3]. Data Mining [4] refers to extracting or mining knowledge from large amounts of data. The aim of data mining is to make sense of large amounts of mostly unsupervised data, in some domain. Classification [5] maps data into predefined groups. It is often referred to as supervised learning as the classes are determined prior to examining the data. Classification Algorithms usually require that the classes be defined based on the data attribute values.

Literature Survey

Analysis of Related Works

[1] http://www.emedicinehealth.com/diabetes.

[2] http://en.wikipedia.org/wiki/Diabetes_mellitus.

[3] http://diabetes.co.in.

[4] Margaret H. Dunham,-"Data Mining Techniques and Algorithms", Prentice Hall Publishers.

[5] Pardha Repalli, "Prediction on Diabetes Using Data mining Approach".

[6] Joseph L. Breault., "Data Mining Diabetic Databases:Are Rough Sets a Useful Addition? " .

[7] G. Parthiban, A. Rajesh, S.K.Srivatsa, "Diagnosis of Heart Disease for Diabetic Patients using Naive Bayes Method ", International Journal of Computer Applications (0975 – 8887) Volume 24– No.3, June 2011.

[8] P. Padmaja, "Characteristic evaluation of diabetes data using clustering techniques", IJCSNS International Journal of Computer Science and Network Security, VOL.8 No.11, November 2008.

[9] Rajesh.K," Application of Data Mining Methods and Techniques for Diabetes Diagnosis", International Journal of Engineering and Innovative Technology (IJEIT) Volume 2, Issue 3, September 2012

[10] Knowledge Discovery in Databases, http://www2.cs.uregina.ca.

[11] ID3 Algorithm Description, http://en.wikipedia.org /wiki/ID3_algorithm.

Existing System

A decision tree is a classifier expressed as a recursive partition of the instance space. The decision tree consists of nodes that form a rooted tree meaning it is a directed tree with a node called "root" that has no incoming edges. All

ISSN:2319-7900

other nodes have exactly one incoming edge. A node with outgoing edges is called an internal or test node. All other nodes are called leaves (also known as terminal or decision nodes). In a decision tree, each internal node splits the instance space into two or more sub-spaces according to a certain discrete function of the input attributes values.

In the simplest and most frequent case, each test considers a single attribute, such that the instance space is partitioned according to the attribute's value. In the case of numeric attributes, the condition refers to a range. Each leaf is assigned to one class representing the most appropriate target value. Alternatively, the leaf may hold a probability vector indicating the probability of the target attribute having a certain value.

Instances are classified by navigating them from the root of the tree down to a leaf, according to the outcome of the tests along the path. Internal nodes are represented as circles, whereas leaves are denoted as triangles. Note that this decision tree incorporates both nominal and numeric attributes. Each node is labeled with the attribute it tests, and its branches are labeled with its corresponding values.

Proposed System

ID3 Algorithm

The basic idea of ID3 algorithm is to construct the decision tree by employing a top-down, greedy search through the given sets to test each attribute at every tree node. In order to select the attribute that is most useful for classifying a given sets, we introduce a metric - information gain.

To find an optimal way to classify a learning set, what we need to do is to minimize the questions asked (i.e. minimizing the depth of the tree). Thus, we need some function which can measure which questions provide the most balanced splitting. The information gain metric is such a function.

The first term in the equation for *Gain* is just the entropy of the original collection *S* and the second term is the expected value of the entropy after S is partitioned using attribute *A*. The expected entropy described by this second term is simply the sum of the entropies of each subset, weighted by the fraction of examples $\frac{|S_v|}{|S|}$ that belong to *Gain (S, A)* is therefore the expected reduction in entropy caused by knowing the value of attribute *A*.

The process of selecting a new attribute and partitioning the training examples is now repeated for each non terminal descendant node. Attributes that have been incorporated higher in the tree are excluded, so that any given attribute can appear at most once along any path through the tree. This process continues for each new leaf node until either of two conditions is met:

1. Every attribute has already been included along this path through the tree, or
2. The training examples associated with this leaf node all have the same target attribute value (i.e., their entropy is zero).

A. Advantages:

- Reduce the number of nodes while constructing the tree.
- Speed is high and memory usage is very low.
- Using bottom up strategy, the nodes are searched.
- Apply Information Gain & Entropy to get the accurate results.

B.Result Analysis:

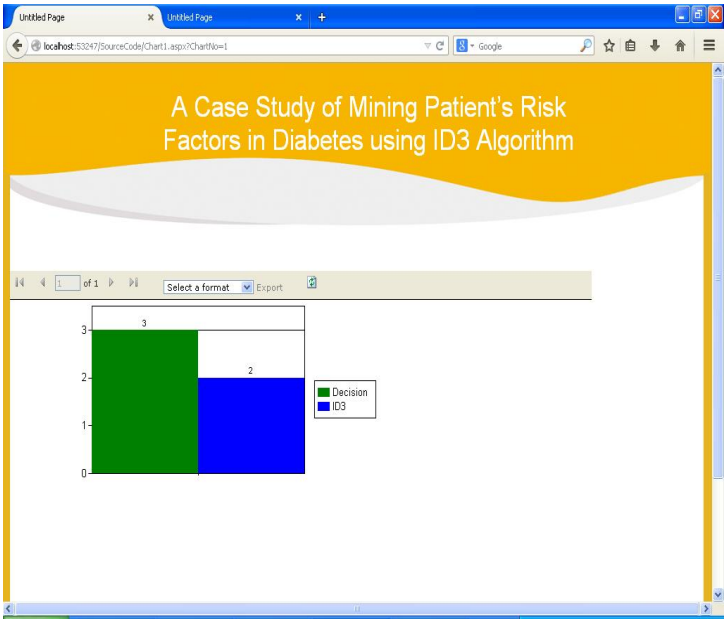

Figure 1.Memory Usage Comparison (Decision Tree & ID3)

By comparing the Decision Tree and ID3, ID3 algorithm produces the best result in less time and less memory space.

ISSN:2319-7900

S.NO.	DataSet Name	AvgTime in Milli Seconds		Memory Usage in KB	
		Decision Tree	ID3	Decision Tree	ID3
1	City1– Anthiyur	480	101	20.43	11.12
2	City2 – Bhavani	360	190	26.27	10.05
3	City3 - Ammapet	590	216	18.26	8.23

Table 1. Average Time and Memory Usage – Comparison

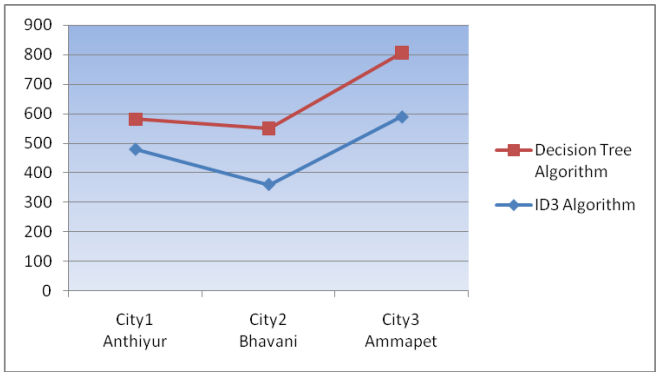

Figure 2. Average Performance Time Comparison

Figure 3. Memory Usage Comparison

CONCLUSION AND FUTURE WORK

A. CONCLUSION

In this paper, a case study has been made with the information like causes, age of the patient, gender, symptoms, risk ratio of the patients and end of the test, to predict major causes on the diabetes for patients. As there are many approaches that are used for data classification, the decision tree method (ID3 algorithm) is used here in evaluating major causes.

In future, we want to study the effect of analyzing the causes from the patients and the entropy values have been calculated by reducing the number of nodes. We want to make the node search in a quicker manner.

B. SCOPE FOR FUTURE WORK

The future involves the in analyzing the dataset by using the advanced algorithm such as ASSISTANT, C4.5. It reduces the node and makes the performance faster. It includes removing the repeated nodes, constructing the tree with minimum nodes, partitioning the data items based on the causes.

The survey is based on the Erode district and from that various cities are analyzed and the impact factor in processed. In future, more number of district comparison may takes place and analysis may be done. The future work will reduce the redundant data, improves the performance, effectiveness and accuracy of the search. This extension will give the better result for the performance ratio.

REFERENCES

[1] http://www.emedicinehealth.com/diabetes.

[2] http://en.wikipedia.org/wiki/Diabetesmellitus.

[3] http://diabetes.co.in.

[4] Margaret H. Dunham,-"Data Mining Techniques and Algorithms", Prentice Hall Publishers.

[5] Pardha Repalli, "Prediction on Diabetes Using Data mining Approach".

[6] Joseph L. Breault., "Data Mining Diabetic Databases:Are Rough Sets a Useful Addition? " .

[7] G. Parthiban, A. Rajesh, S.K.Srivatsa, "Diagnosis of Heart Disease for Diabetic Patients using Naive Bayes Method ", International Journal of Computer Applications (0975 – 8887) Volume 24– No.3, June 2011.

[8] P. Padmaja, "Characteristic evaluation of diabetes data using clustering techniques", IJCSNS International Journal of Computer Science and Network Security, VOL.8 No.11, November 2008.

[9] Rajesh.K," Application of Data Mining Methods and Techniques for Diabetes Diagnosis", International Journal of Engineering and Innovative Technology (IJEIT) Volume 2, Issue 3, September 2012

[10] Knowledge Discovery in Databases, http://www2.cs.uregina.ca.

[11] ID3 Algorithm Description, http://en.wikipedia.org /wiki/ID3_algorithm.

[12] Gryzmala-Busse, Jery W. "Selected Algorithms of Machine Learning from Examples".

[13]Guyon, U.V. Luxburg, and R.C. Williamson, "Clustering: Science or Art?," Proc. NIPS Workshop Clustering Theory, 2009.

ISSN:2319-7900

[14] A. Banerjee, S. Merugu, I. Dhillon, and J. Ghosh, "Clustering with Bregman Divergences," J. Machine Learning Research, vol. 6, pp. 1705-1749, Oct. 2005.
[15] Quinlan, J. R. 1986. Induction of Decision Trees. Mach. Learn. 1, 1 (Mar. 1986), 81-106.
[16] Grzymala-Busse, Jerzy W. "Selected Algorithms of Machine Learning from Examples." *Fundamenta Informaticae* 18, (1993): 193–207.
[17] Mitchell, Tom M. *Machine Learning*. McGraw-Hill, 1997. pp. 55–58.
[18]www.cs.uregina.ca
[19] www.cise.uif.edu
[20]www.cis.temple.edu