Journal of Computer Science 10 (7): 1151-1155, 2014
ISSN: 1549-3636

doi:10.3844/jcssp.2014.1151.1155 Published Online 10 (7) 2014 (http://www.thescipub.com/jcs.toc)

ITERATIVE DICHOTOMISER-3 ALGORITHM IN DATA MINING APPLIED TO DIABETES DATABASE

P. Vasudevan

Department of Computer Science and Engineering,
Mookambigai College of Engineering, Anna University, Chennai, India

Received 2013-07-27; Revised 2013-12-13; Accepted 2014-02-15

ABSTRACT

In this study, eight major factors playing significant role in the Pima Indian population are analyzed. Real time data is taken from the large dataset of National Institute of Diabetes and Digestive and Kidney Diseases. The data is subjected to an analysis by logistic regression method using SPSS 7.5 statistical software, to isolate the most significant factors among the eight factors taken. Then the significant factors are further applied to decision tree technique called the Iterative Dichotomiser-3 algorithm which leads to significant conclusions about this diabetes disorder which poses to be a greatest threat to mankind in the coming era. Conglomeration of data mining techniques and medical data base research can lead to life saving conclusions for the physicians at critical times to save the mankind.

Keywords: BMI, Diabetes, Decision Tree, Logistic Regression, Plasma

1. INTRODUCTION

ID3 begins by choosing a random subset of the training instances. This subset is called the window. The procedure builds a decision tree that correctly classifies all instances in the window. The tree is then tested on the training instances outside the window. If all the instances are classified correctly then the procedure halts. Otherwise it adds some of the instances incorrectly classified to the window and repeats the process. This iterative strategy is empirically more efficient than considering all instances at once. In building a decision tree ID3 selects the feature which minimizes the entropy function given below and thus best discriminates among the training instances. Data have been collected from about 768 Indian Origin females who were tested for the presence of diabetes mellitus of which 268 were found to be positive. Sample of 336 records are selected deleting the record sets with zero values.

1.1. Decision Trees in Data Mining (Quinlan, 1986)

In data mining, a decision tree is a predictive model; that is, a mapping of observations about an item to conclusions about the item's target value. More descriptive names for such tree models are classification tree or reduction tree. In these tree structures, leaves represent (Ankerst *et al*., 1999) classifications and branches represent conjunctions of features that lead to those classifications. The machine learning technique for inducing a decision tree from data is called decision tree learning, or decision trees. In decision theory and decision analysis, a decision tree is a graph or model of decisions and their possible consequences, including chance event outcomes, resource costs and utility. It can be used to create a plan to reach a goal. Decision trees are constructed in order to help with making decisions (Almuallim, 1996). A decision tree is a special form of tree structure and a descriptive means for calculating conditional probabilities.

Decision tree learning is a common method used in data mining. Each interior node corresponds to a variable; an arc to a child represents a possible value of that variable. A leaf represents a possible value of target variable given the values of the variables represented by the path from the root. A tree can be “learned” by splitting the source set into subsets based on an attribute value test. This process is repeated on each derived subset in a recursive manner. The recursion is completed when splitting is either non-feasible, or a singular classification can be applied to each element of the

derived subset. A random forest classifier uses a number of decision trees, in order to improve the classification rate. In data mining (Brodley and Utgoff, 1995), trees can be described also as the synergy of mathematical and computing techniques that aids on the description, categorization and generalization of a given set of data. Data comes in records of the form:

$$(X_i, y) = (x_1, x_2, x_3 ... x_k, y)$$

The dependent variable, y, is the variable that we are trying to understand, classify or generalize. The other variables x_1, x_2, x_3 are the variables that will help us for predictions.

1.2. Building Decision Trees (Quinlan, 1993)

- Top-down tree construction
- Bottom-up tree pruning

1.3. Choosing the Splitting Attribute

- At each node, available attributes are evaluated on the basis of separating the classes of the training examples. A Goodness function is used for this purpose
- Typical goodness functions
- Information Gain (ID3)
- Information Gain Ratio
- Gini Index

1.4. Information Entropy by Claude Shannon

In information theory, the Shannon entropy (Ankerst *et al*., 1999) or information entropy is a measure of the uncertainty associated with a random variable. It can be interpreted as the average shortest message length, in bits, that can be sent to communicate the true value of the random variable to a recipient. This represents a fundamental mathematical limit on the best possible lossless data compression of any communication: The shortest average number of bits that can be sent to communicate one message out of all the possibilities is the Shannon entropy (Shannon, 1948),

Formula for computing the entropy:

$$\text{Shannon Entropy } (p_1, p_2, \ldots, p_n) = -p_1\log p_1 - p_2\log p_2 \ldots - p_n\log p_n$$

1.5. Definition of Information Entropy

The information entropy of a discrete random variable X, that can take the range of possible values $\{x_1 ... x_n\}$ is defined to be:

$$H(X) = E\left(I(X)\right) = \sum_{i=1}^{n} p(x_i)\log_2 p(x_i)$$

I(X) is the information content or self-information of X, which is itself a random variable and $p(x_i) = P(X = x_i)$ is the probability mass function of X.

Iterative Dichotomiser-3 (ID3) is an algorithm used to generate a decision tree. However, it does not always produce the smallest tree and is therefore a heuristic. Occam's razor is formalized using the concept of information entropy as:

$$I_E(i) = \sum_{j}^{m} f(i, j)\log f(i, j)$$

2. LOGISTIC REGRESSION OUTPUTS FROM SPSS 7.5

Logistic regression method (Tsien *et al*., 1998) was applied to bring out the significance factors like age, obesity in the cause of the diabetes disorder, in Pima Indian diabetes database using SPSS 7.5 software. These factors are fuzzified to form a sample decision tree by ID3 algorithm.

2.1. Analysis of Logistic Regression Results

Total number of cases (Unweighted):	336
Number of selected cases:	336
Number of unselected cases:	0
Number of selected cases:	336
Number rejected because of missing data:	0
Number of cases included in the analysis:	336

After dependent variable Encoding and from the observations of **Table 1**, we find that the following factors playing significant role in the cause of diabetes:

- Age: Sig = 0.0337 so 97% confidence level
- Body Mass Index: Sig = 0.0164 = 0.02 so 98% confidence level
- PDF (Diabetes Pedigree Function): Sig = 0.0216 so 98% confidence level. Implication of Hereditary Nature in the disease
- Plasma (Glucose Concentration in Saliva): Sig = 0.0000 100% confidence level as shown in **Fig. 1**

Table 1. Logistic regression outputs from SPSS 7.5

	---------- Variables in the equation ----------						
Variable	B	S.E.	Wald	df	Sig	R	Exp (B)
AGE	0.0408	0.0192	4.5086	1	0.0337	0.0767	1.0416
BMI	0.0761	0.0315	5.7548	1	0.0164	0.0938	1.0791
BP	0.0060	0.0132	0.2063	1	0.6497	0.0000	1.0060
INSU	0.23E-05	0.0014	0.0005	1	0.9822	0.0000	1.0000
PDF	1.0970	0.4776	5.2746	1	0.0216	0.0876	2.9951
PLAS	0.0362	0.0062	33.4697	1	0.0000	0.2717	1.0368
PRG	0.0736	0.0597	1.5197	1	0.2177	0.0000	1.0764
THICK	0.0111	0.0187	0.3525	1	0.5527	0.0000	1.0112
Constant	-10.8272	1.4227	57.9161	1	0.0000		

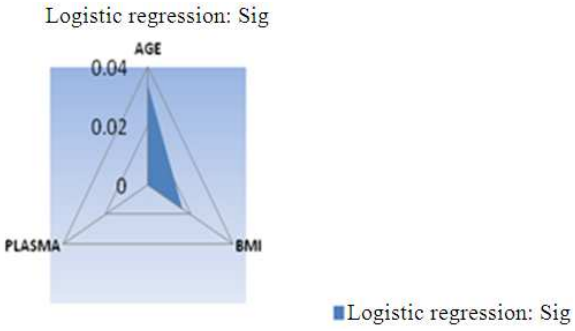

Fig. 1. Significant Factors from Logistic Regression output

2.2. The ID3 Algorithm Applied to Diabetes Database (Almuallim, 1996)

Select a random subset W (called the "window") from the training set.

Build a decision tree for the current window. Select the best feature which minimizes the entropy function H.

$H = \sum -p_i \log p_i$ (optimal values are available and the optimum entropy may be found by discrete probabilistic methods).

Where p_i is the probability associated with i^{th} class. The entropy is calculated for each value. The sum of the entropy is calculated for each value. The sum of the entropy weighted by the probability of each value is the entropy for the feature. Categorize training instances into subsets by this feature. Repeat this process recursively until each subset contains instances of one kind (class) or some statistical criterion is satisfied.

Scan the entire training set for exceptions to the decision tree.

If exceptions are found, insert some of them into W and repeat from Step 2. The insertion may be done either by replacing some of the existing instances in the window or by augmenting it with the new exceptions. In practice a statistical criterion can be applied to stop the tree from growing as long as most of the instances are classified correctly **Fig. 2**.

2.3. Pesudo Code for ID3 Algorithm (Grefenstette *et al*., 1990)

```
function ID3 (I, 0, T) {
/*I is the set of input attributes
*O is the output attribute
*T is a set of training data
**function ID3 returns a decision tree*/
if (T is empty)
{
return a single node with the value
 "wrong";
}
if (all records in T have the same value for O) {
return a single node with that value;
}
if (I is empty)
{
return a single node with the value of the most frequent
value of O in T;
          */
}
/* case where we can't return a single node */
compute the information gain for each attribute in I
relative to T; let X be the attribute with largest Gain(X,
T) of the attributes in I;
 Let {x_j| j = 1,2, .., m} be the values of X;
Let {T_j| j = 1,2, .., m} be the subsets of T when T is
partitioned according the value of X;
return a tree with the root node labelled X and arcs
labelled x1, x2, x3... xm, where the arcs go to
 the trees ID3(I-{X}, O, T_1), ID3(I-{X}, O, T_2)...
ID3(I-{X}, O, T_m);
 }
```

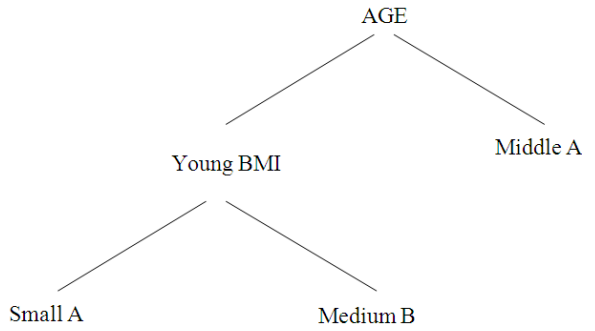

Fig. 2. Sample of Decision Tree by ID3 Algorithm

Table 2. Sample Fuzzified Dataset from Pima Indian diabetes database: (AIP, 2014)

CLASS	BMI	AGE	PLASMA
A	Small	Young	Less
A	Medium	Middle	Less
A	Medium	Middle	Less
A	Big	Middle	High
B	Medium	Young	Less
B	Medium	Young	Less

2.4. Application of ID3 Algorithm in Samples of Pima Indian Diabetes Database

The ID3 algorithm builds a decision tree for classifying the following (Nunez, 1991) Objects: Sample Data of Pima Indian Diabetes Database Class A: Acquired Diabetes Class B: Non Diabetic.

First, we calculate the entropy for each attribute.

BMI H:

= 1/6(-1/1 log1/1) + 4/6(-2/4 log2/4 2/4log2/4) +1/6(-1/1log1/1)
= 0.462

AGE H:

= 3/6(-2/3log2/3-1/3log1/3) + 3/6(-3/3 log3/3)
= 0.318

PLASMA H:

= 5/6(-3/5log3/5 - 2/5 log2/5) + 1/6(-1/1log1/1)
= 0.56

Thus from **Table 2** we select the attribute AGE as the first decision node since it is associated with the minimum entropy. This node has two branches: Young and middle. Under the branch middle, only class a objects fall and hence no further discrimination is needed. Under the branch young, we need another attribute to make further distinctions. So, we calculate the entropy for the other two attributes under this branch (Quinlan, 1987).

BMI H:

=1/3 (-1/1log1/1) + 2/3(-2/2log2/2)
= 0

PLASMA H:

= 3/3(-2/3 log2/3 - 1/3 log1/3)
= 0.636

We use the attribute BMI as the second decision node which has a minimal value.

3. CONCLUSION

Data mining method using logistic regression implies that Age, Obesity, PDF and Plasma level are to be taken care of for the onset of diabetes mellitus. ID3 algorithm applied to the sample database gives the decision tree prediction with major factors of diabetes. The paper on a small scale tries to bring out the dominant factors alone by applying Iterative Dichotomiser ID3 algorithm of data mining. As our mankind has a great threat of this pancreatic disorder more in the coming era the sample data is chosen from diabetes database. The same idea can be applied to any disease database on a large sampling to bring out more useful diagnostic findings before complications affect the human population.

4. REFERENCES

AIP, 2014. Home page for the National Institute of diabetes and digestive and kidney diseases. AIP Publishing LLC.

Almuallim, H., 1996. An efficient algorithm for optimal pruning of decision trees. Artif. Intell., 83: 347-362. DOI: 10.1016/0004-3702(95)00060-7

Ankerst, M., C. Elsen, M. Ester and H.P. Kriegel, 1999. Visual classification: An interactive approach to decision tree construction. Proceedings of the 5th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Aug. 15-18, ACM Press, New York, USA., pp: 392-396. DOI: 10.1145/312129.312298

Brodley, C.E. and P.E. Utgoff, 1995. Multivariate decision trees. Mach. Learn., 19: 45-77. DOI: 10.1023/A:1022607123649

Grefenstette, J.J., C.L. Ramsey and A.C. Schultz, 1990. Learning sequential decision rules using simulation models and competition. Mach. Learn., 5: 355-381. DOI: 10.1007/BF00116876

Nunez, M., 1991. The use of background knowledge in decision tree induction. Mach. Learn., 6: 231-250. DOI: 10.1007/BF00114778

Quinlan, J.R., 1986. Induction of decision trees. Mach. Learn., 1: 81-106. DOI: 10.1023/A:1022643204877

Quinlan, J.R., 1987. Simplifying decision trees. Int. J. Man Mach. Stud., 27: 221-234. DOI: 10.1016/S0020-7373(87)80053-6

Quinlan, J.R., 1993. C4.5: Programs for Machine Learning. 1st Edn., Morgan Kaufmann, San Mateo, ISBN-10: 1558602380, pp: 302.

Shannon, C.E., 1948. A mathematical theory of Shannon communication. Bell Syst. Tech. J., 27: 379-423. DOI: 10.1002/j.1538-7305.1948.tb01338.x

Tsien, C.L., H.S. Fraser, W.J. Long and R.L. Kennedy, 1998. Using classification tree and logistic regression methods to diagnose myocardial infarction. Stud. Health Technol. Inform., 52: 493-497. PMID: 10384505