ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 8, August 2014

Prediction of Missing Values in Blood Cancer & Occurrence of Cancer Using Improved Id3 Algorithm

Priyadharsini.C[1], Dr. Antony Selvadoss Thanamani[2]

M.Phil, Department of Computer Science, NGM College, Pollachi, Coimbatore, India[1]

HOD, Department of Computer Science, NGM College, Pollachi, Coimbatore, India[2]

ABSTRACT: Missing data can be recreant because it is complicated to identify the problem. Missing data can cause critical problems. First, most statistical procedures automatically eliminate cases with missing data. . Second, the analysis might run but the results may not be statistically significant because of the small amount of input data. In this paper we inspect the enforcement of two unusual data imputation process in a task where the aim is to conclude the probability of finding missing data in blood cancer and occurrence of blood cancer using improved ID3 algorithm. Cancer is one of the deadliest diseases found among many people across the world. Our project aims at helping the medical practitioners to diagnose the patients at the early stage which can reduce the number of deaths.

KEYWORDS: Data mining, missing values, ID3 Algorithm, data migration, decision tree classification, multi array model, data density clustering.

I INTRODUCTION

To find out the missing values, sometimes prediction may use to fill the data. Prediction should be highly accurate. So present we are executing a Multidimensional Array model with modified ID3 algorithm. The modified ID3 algorithm will compare the current spatial database with the normal database from the input database. From the data set, an operational database will be created for the cancer patients and a database for normal patients. This database will be individual and many numbers of practical data are available. The result will be recovered from the dataset. Whether the patient will affect from cancer or not, also their infection ration percentage can be find out, Along with the lost values in the database during the time of data migrating. And also to analyzing the occurrences of cancer patients using data density clustering.

II METHODOLOGY

a) Attribute Selection in multi array model

How does ID3 decide which attribute multi array model is the best. A statistical property, called information gain, is used. Gain magnitude how well a given attribute split up training examples into targeted classes. The one with the highest information (information being the most useful for classification) is selected. In order to assign gain, we first borrow an idea from information theory called entropy. Entropy calculates the amount of information in an attribute.

Given a collection S of c outcomes in multiarray

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 8, August 2014

Entropy(S) = S -p(I) log2 p(I)

Where p(I) is the proportion of S belonging to class I. S is over c. Log2 is log base 2.

Note that S is not an attribute but the entire sample set.

b) Characters of ID3 algorithm in multi array model

Detailed elaborations are presented for the idea on ID3 algorithm of Decision Tree. An improved method called Improved ID3 algorithm that can improve the speed of generation is brought forward owing to the disadvantages of ID3 algorithm. Moreover, based on Improved ID3 algorithm, data mining for Blood-cancers is carried out for principallyanticipating the communication between recurrence and other attributes of breast cancer by building use of SQL Server 2005 Analysis Services. Results prove the strength of Decision Tree in medical data mining which provide physicians with diagnostic assistance. The basic convention of decision tree for constructing tree can be depicted by ID3 algorithm. It uses the divide-and-conquer strategy in the development of decision tree, which uses the information gain of characteristic as the examining function of attribute selection of a branch ineach node of the tree, selecting the information gain as the characteristic of the branch.

c) ID3 in multi array model algorithm is described as follows

Let E = D1× D2× ... × Dn be finite-dimensional vector n, where Dj

is a finite set of discrete symbols, E elements e = <v1, v2, ... , vn> is the sample, vjDj, j = 1, 2, ..., n. Let PE be the positive sample set, NE be the anti-sample set, and the number of samples which are p and n. rendering to the regulations of information theory.

d) ID3 algorithm is based on two assumptions:

(1) In the vector space E, a decision tree classification probability for any sample and the probability for positive sample and anti-sample in E are the same.

(2) The expected bits of information needed for making the correct identification by a decision tree are:

If attribute A is the root of the decision tree, A has n values {u1, u2, ... , un}, which will divide the sample set E into n subsets {E1, E2, ... , En}. Supposing that Eicontains pi positive samples and negative samples, then a subset of the information needed for the Eiis I (pi+ ni), and the expected information needed for the attribute A as the root node.Therefore, the information gain of classification attribute ofA as the root node is Gain (A) = I (p, n)-E (A). ID3 algorithm selection contributes the greatest attribute of Gain (A) to a branch of the node attributes, and each node of the decision tree is using this principle until the decision tree is completed (each node of the samples belong to the same class or all Category attributes are used up). One advantage of ID3 is its time of tree construction and difficulty of the task (such as the number of sample set samples, the number of attributes for each sample to study the complexity of the concept of the decision tree nodes) are steadily increasing in linear and the computation is relatively small.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 8, August 2014

III LITERATURE SURVEY

1. A Literature Review of Data Mining Techniques used in Healthcare Databases.

In this paper we present an overview of the current research being carried out using the data mining techniques for the diagnosis and prognosis of various diseases. The goal of this study is to identify the most well performing data mining algorithms used on medical databases. The following algorithms have been identified: Decision Trees, Support Vector Machine, Artificial neural networks and their Multilayer Perception model, Naïve Bayes, Fuzzy Rules. Analyses show that it is very difficult to name a single data mining algorithm as the most suitable for the diagnosis and/or prognosis of diseases

2. Comparing Performance of logistic regression, decision tree and neural network for classifying heart disease patients

In this study, performances of classification techniques were compared in order to predict the presence of the patients getting a heart disease. A retrospective analysis was performed in 303 subjects. We compared the performance of logistic regression (LR), decision trees (DTs), and Artificial neural networks (ANNs). The variables were medical profiles are age, Sex, Chest Pain Type, Blood Pressure, Cholesterol, Fasting Blood Sugar, Resting ECG, Maximum Heart Rate, Induced Angina, Ole Peak, Slope, Number Colored Vessels, Thal and Concept Class. We have created the model using logistic regression classifiers, artificial neural networks and decision trees that they are often used for classification problems. Performances of classification techniques were compared using lift chart and error rates.

3. Support System for Medical Diagnosis Using Data Mining

Knowledge discovery from the dramatically increased data of an auto-stored medical information system is still in its infancy. The purpose of this study is to use widely available and easily operated techniques that can satisfy general users in extracting specific knowledge to make the medical information system more functional. Data mining techniques, including data visualization, correlation analysis, Discriminant analysis, and neural networks supervised classification, were applied to heart disease databases. These techniques can help to identify high risk patients, define the most important factors (variables) in heart disease, and build a multivariate relationship model to show the relationship between any two variables in a way that such relationships are easy to view.

4.Data Mining in Health care: Current Applications and Issues.

The successful application of data mining in highly visible fields like e-business, marketing and retail have led to the popularity of its use in knowledge discovery in databases (KDD) in other industries and sectors. Among these sectors that are just discovering data mining are the fields of medicine and public health. This research paper provides a survey of current techniques of KDD, using data mining tools for healthcare and public health. It also discusses critical issues and challenges associated with data mining and healthcare in general. The research found a growing number of data mining applications, including analysis of health care centers for better health policy-making, detection of disease outbreaks and preventable hospital deaths, and detection of fraudulent insurance claims.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 8, August 2014

5. Application of ID3 Algorithm in Information Asset Identification

In the issue of information security risk evaluation, the asset, the threat and the vulnerability are the three most important elements. Information asset identification is a primary link of information security risk evaluation process. In this paper, the decision tree algorithm was applied into the identification of information assets; the basic process of ID3 algorithm are described; the data of information system property is classified by ID3 algorithm; the decision tree was made; and the rules are extracted for providing basis of information asset recognition. Information security risk evaluation is the basis for an organization to ensure information security. Among the factors that information security risk evaluation involved, the information asset is the foremost important one. Information assets are the primary object of the information system security policy.

IV PROJECT PROPOSAL

The project proposal is the attempt to respond to or take benefit of a particular situation and is an essential ingredient for correctly launching the system analysis. Although there are no hard and fast rules as to the form and content of the project proposal, the proposal should address the following points:

- The specifics of the business situation or problem.
- The significance of the problem to the organization.
- Alternative solutions.
- The possible use of computer information systems to solve the problem.

The various people interested in or possessing knowledge relevant to the problem.

System projects that are to be shared by a number of departments and users are usually approved by a committee rather than an individual. A project proposal is submitted to a committee that determines the merits of the proposal and decides whether or not to approve it. The board is made up of people from various functional areas of the organization that have acuriosity in the operation and information of the proposed system.

V PROPOSED SYSTEM

Proposed Method: Decision tree classification method using modified ID3 algorithm.

Cancer is one of the deadliest diseases found among many people across the world. Our project aims at helping the medical practitioners to diagnose the patients at the early stage which can reduce the number of deaths. The decision tree is an important classification method in data mining classification. The proposed work is that we have modified the id3 algorithm using decision tree classification method and included the pre processing steps for the cancer data set to improve the accuracy of the classifier. The data set has missing values in it. In the pre processing steps of the data set, we have resolved it. Also the data set has data conflicts in it. And we have proposed an approach to resolve it. The concept of conditional entropy measure is used in the id3 algorithm and modified it. Then after pre processing the data set, it is supplied to the modified algorithm which constructs the decision tree and thus it proves to increase the accuracy of the classifier.

The proposed sample data used by ID3 has certain requirements, which are:

Attribute-value description - the same attributes must describe each example and have a fixed number of values.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 8, August 2014

Predefined classes - an example's attributes must already be defined, that is, they are not learned by ID3.

Discrete classes - classes must be sharply described. Continuous classes broken up into vague categories such as a metal being "hard, quite hard, flexible, soft, quite soft" are suspect.

Sufficient examples - since inductive generalization is used (i.e. not provable) there must be enough test cases to distinguish valid patterns from chance occurrences.

Decision tree learning algorithm has been successfully used in expert systems in capturing knowledge. The main task performed in these systems is using inductive methods to the given values of attributes of an unknown object to determine appropriate classification according to decision tree rules. We examine the decision tree learning algorithm ID3 and implement this algorithm using C# programming. We first implement basic ID3 in which we dealt with the target function that has discrete output values. We also extend the domain of ID3 to real valued output, such as numeric data and discrete outcome rather than simply Boolean value.

a) Improved Algorithm Description

Input:

Examples: A data set consisting of the training tuples.

Target _ Attribute, associated target values for the above given tuples.

Attributes, names of the attributes in the dataset.

Output:

T, a decision tree with its corresponding decision rules.

Method:

ID3 (Example, Target_Attribute, Attributes)

Create a root node for the tree

If all Example are positive, Return the single-node tree Root, with label = +.

If all Example are negative, Return the single-node tree Root, with label = -.

If number of predicting attributes is empty, then Return the single node tree Root,

with label = most common value of the target attribute in the Example.

Otherwise Begin

A = The Attribute that best classifies Example.

Decision Tree attribute for Root = A.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 8, August 2014

For each possible value, vi, of A,

Add a new tree branch below Root, corresponding to the test A = vi.

Let Example(vi) be the subset of Example that have the value vi for A

If Example(vi) is empty

Then below this new branch add a leaf node with label = most

common target value in the Example

Else below this new branch add the subtree ID3 (Example(vi),

Target_Attribute, Attributes – {A})

End

Return Root

b) Improvised ID3 Algorithm

The classical ID3 algorithm is firstly applied with importance of each attribute. Then, information gain is combined with attribute importance, and it is used as a new standard of attribute selection to construct decision tree.

Suppose A is an attribute of data set D, and C is the category attribute of D. the relation

degree function between A and C can be expressed as follows:

$n\ AF\ (A) = \sum | xi1 – xi2 |i=1n$

Where ij x (j = 1, 2 represents two kinds of cases) indicates that attribute A of D takes the

i-th value and category attribute C takes the sample number of the j-th value, n is the number

of values attribute A takes.

Then, the normalization of relation degree function value is followed. Suppose that there

are m attributes and each attribute relation degree function value are AF(1), AF(2), … ,

AF(m) , respectively. Thus, there is

$V\ (k) = AF\ (k)$, where $0 < k \leq m$.

$AF\ (1) + AF\ (2) + \ldots + AF\ (m)$

Thus, the improvised version of Information Gain formula becomes:

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 8, August 2014

Gain` (A) = (I(S1, S2, … ,Sm) – E(A)) * V(A) * (n-m).

VI CONCLUSION

The input data set will be measured as a data array with multiple dimension functionality. This is to increase the accuracy of finding the missing data. An improved ID3 algorithm is to find the missing values in the cancer patient dataset. In existing Method Decision tree classification algorithm is used. The proposed method Decision tree classification method using modified ID3 algorithm is used.The proposed model is simple to understand and interpret, requires little data preparation, Able to handle both numerical and categorical data Uses a white box model, possible to validate a model using statistical tests, Robust, Performs well with large datasets.

REFERENCES

[1] QuKaishe, Cheng Wenli, Wang Junwang. Improved Algorithm Based on ID3[J]. Computer Engineering and Applications. 39(25): 104107, 2003.
[2] Jiawei Han and MichelineKamber, "Data Mining: Concepts and Techniques", Second Edition
[3] Huang Ming1, NiuWenying1, Liang Xu ,"An improved decision tree classification algorithm based on ID3 and the application in score analysis", Chinese Control and Decision Conference (CCDC 2009).
[4] Chen Jin, Luo De-lin, Mu Fen-xiang, "An Improved ID3 Decision Tree Algorithm", Proceedings of 4th International Conference on Computer Science & Education 2009
[5] NishantMathur, Sumit Kumar, Santosh Kumar, andRajni Jindal ,"The Base Strategy for ID3 Algorithm of Data Mining Using Havrda and Charvat Entropy Based on Decision Tree" ,International Journal of Information and Electronics Engineering, Vol. 2, No. 2, March 2012.
[6] L.Sathish Kumar, Mrs.A.Padmapriya,"ID3 Algorithm Performance of Diagnosis For Common Disease", International Journal of Advanced Research in Computer Science and Software Engineering Volume 2, Issue 5, May 2012.
[7] Karmaker et al."Incorporating an EM-Approach for Handling Missing AttributeValues in Decision Tree Induction"
[8] Wai-Ho Au, Member, IEEE, Keith C. C. Chan, Andrew K. C. Wong, Fellow, IEEE,and Yang Wang, Member, IEEE "Attribute Clustering for Grouping, Selection, and Classification of Gene Expression Data", Manuscript received Sep. 15, 2004; revised Dec. 1, 2004; accepted March 1, 2005. The work by W.-H. Au and K. C. C. Chan were supported in part by The Hong Kong Polytechnic University under Grants A-P209 and G-V958.

BIOGRAPHY

C.Priyadharsiniis a Research scholar in Department of Computer Science, Nallamuthu Gounder Mahalingam College,Pollachi. She received her Master of Science in Software systemsM.Sc (ss) from Sree Saraswathi Thyagaraja College, Pollachi under Bharathiar University, Coimbatore. She has presented a paper in National level conference and attended Workshop / Seminar. Her research focuses on Data Mining (missing data imputation).

Dr. Antony SelvadossThanamani is presently working as Professor and Head, Dept of Computer Science, NGM College, Coimbatore, India (affiliated to Bharathiar University, Coimbatore). He has published more than 100 papers in international/ national journals and conferences. He has authored many books on recent trends in Information Technology. His areas of interest include E-Learning, Knowledge Management, Data Mining, Networking, Parallel and Distributed Computing. He has to his credit 24 years of teaching and research experience. He is a senior member of International Association of Computer Science and Information Technology, Singapore and Active member of Computer Science Society of India, Computer Science Teachers Association,New York.