ISSN 2229-5518

PREVALANCE AND RISK FACTORS OF ANEMIA ALONG WITH CLASSIFIERS

[1] P.Usha [2] S.Saranya [3] N.Tamilselvi
Assistant professor
Department of Computer Science
Dr.NGP Arts and Science College

blessed.usha@gmail.com ssarnyamphil@gmail.com cutenptamil@yahoo.com

Abstract

Healthcare system becomes very important to develop an automated tool that is capable of identifying relevant healthcare information. This work focuses on retrieval of updated, accurate and relevant information from Medline datasets using Machine learning approach. In this paper we present an analysis of the prediction of survivability rate of health problem patients using data mining techniques. Health problem is the leading cause of death all over the world in the past ten years. Several researchers are using statistical and data mining tools to help health care professionals in the diagnosis of health disease. Recently, researchers have been investigating the effect of hybridizing more than one technique showing enhanced results in the diagnosis of diseases. However, using data mining techniques to identify a suitable treatment for these disease patients has received less attention. This paper identifies gaps in the research on various health disease diagnosis and treatment and proposes a model to systematically close those gaps to discover if applying data mining techniques to the health disease treatment data can provide as reliable performance as that achieved in diagnosing health problem. We have investigated the various anemic issues using data mining techniques.

Keywords

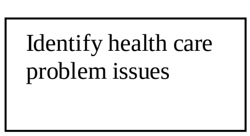
Machine learning algorithm, CHC, CBC, Anemia, Reticindex, RPI.

1. Introduction

It is well known the Information Technology (IT) driven society, knowledge is one of the most significant assets of any organization. The role of IT in health care is well established. Knowledge Management in Health care offers many challenges in creation, dissemination and preservation of health care knowledge using advanced technologies Knowledge discovery in databases is well-defined process consisting of several distinct steps. Data mining is the core step, which results in the discovery of hidden but useful knowledge from massive databases. Health care organizations must have ability to analyze data. Treatment records of millions of patients can be stored and computerized and data mining techniques may help in answering several. This paper presents a process model to guide the data mining projects in the health care sector.

1.1 Knowledge discovery in medical databases

Data mining is an essential step of knowledge discovery. In recent years it has attracted great deal of interest in Information industry. Knowledge discovery process consists of an iterative sequence of data cleaning, data integration, data selection, data mining pattern recognition and knowledge presentation. In particulars, data mining may accomplish class description, association, classification, clustering, prediction and time series analysis. Data mining in contrast to traditional data analysis is discovery driven. Without data mining it is difficult to realize the full potential of data collected within healthcare organization as data under analysis is massive, highly dimensional, distributed and uncertain. Massive healthcare data needs to be converted into information and knowledge, which can help control, cost and maintains high quality of patient care. Healthcare data includes Patient centric data and Aggregate data. For health care organization to succeed they must have the ability to capture, store and analyze data.



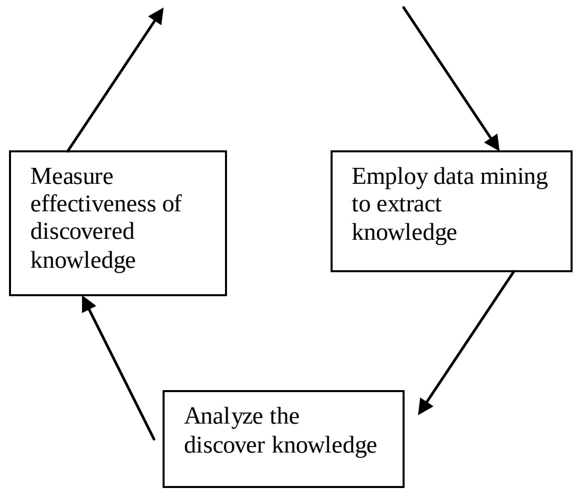


Fig.1: HEALTH ANALYSIS CYCLE

1.2 Data mining techniques in health care

There are various data mining techniques available with their suitability dependent on the domain application. Statistics provide a strong fundamental background for quantification and evaluation of results. However, algorithms based on statistics need to be modified and scaled before they are applied to data mining.

2. METHODOLOGY

2.1 Rule induction

It is the process of extracting useful 'if then' rules from data based on statistical significance. A Rule based system constructs a set of if-then-rules. Knowledge represents has the form IF conditions THEN conclusion this kind of rule consists of two parts. The rule antecedent (the IF part) contains one or more conditions about value of predictor attributes where as the rule consequent (THEN part) contains a prediction about the value of a goal attribute. An accurate prediction of the value of a goal attribute will

ISSN 2229-5518

improve decision-making process. IF-THEN prediction rules are very popular in data mining; they represent discovered knowledge at a high level of abstraction. In the health care system it can be applied as follows:

(Symptoms) (Previous--- history) ------- >
(Cause—of--- disease)

Rule Induction Method has the potential to user retrieved cases for predictions. The following example gives rule induction method for prediction anemic concentration

2.3 ANEMIA

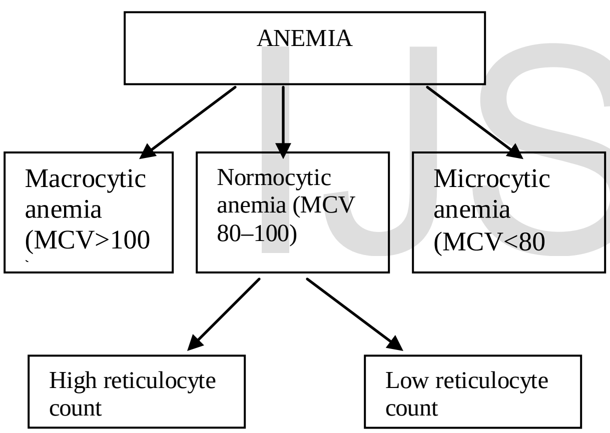


Fig.2: Types of Anemia

2.2 Decision tree

It is a knowledge representation structure consisting of nodes and branches organized in the form of a tree such that, every internal non-leaf node is labeled with values of the attributes. The branches coming out from an internal node are labeled with values of the attributes in that node. Every node is labeled with a class (a value of the goal attribute).Tree based models which include classification and regression trees, are the common implementation of induction modeling. Decision tree models are best suited for data mining. They are inexpensive to construct, easy to interpret, easy to integrate with database system and they have comparable or better accuracy in many applications. There are many Decision tree algorithms such as HUNTS algorithm (this is one of the earliest algorithm), CART, ID3, C4.5 (a later version ID3 algorithm), SLIQ, SPRINT. The decision tree shown in Fig. 2 is built from the very small training set (Table 2). In this table each row corresponds to a patient record. We will refer to a row as a data instance. The data set contains three predictor attributes, namely Age, Gender, Intensity of symptoms and one goal attribute, namely disease whose values (to be predicted from symptoms) indicates whether the corresponding patient have a certain disease or not.

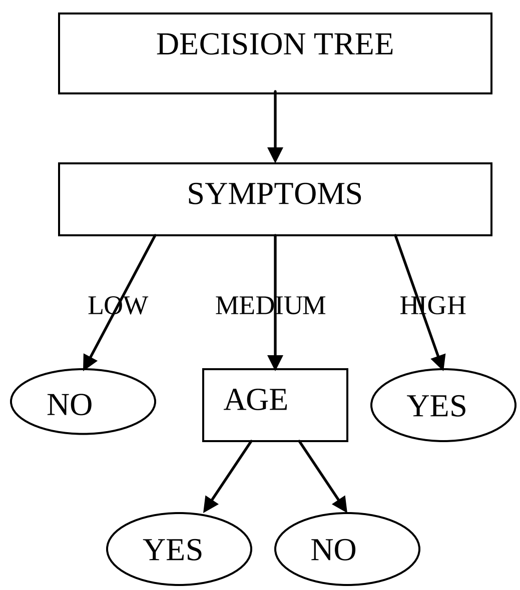


Fig. 2: A decision tree

Attributes list

Age	Gender	Intensity of symptoms	Disease (goal)
20	Male	medium	yes
23	Male	high	yes
22	Female	medium	yes
21	Female	low	no
22	Male	low	no
18	Female	low	no
19	Male	medium	no

Table 1: Data set used to build decision tree

Decision tree can be used to classify an unknown class data instance with the help of the above data set given in the Table 1. The idea is to push the instance down the tree, following the branches whose attributes values match the instances attribute values, until the instance reaches a leaf node, whose class label is then assigned to the instance. For example, the data instance to be classified is described by the tuple (Age=23, Gender=female, Intensity of symptoms = medium, Goal =?), where "?" denotes the unknown value of the goal instance. In this example, Gender attribute is irrelevant to a particular classification task. The tree tests the intensity of symptom value in the instance. If the answer is medium; the instance is pushed down through the corresponding branch and reaches the Age node. Then the tree tests the Age value in the instance. If the answer is 23, the instance is again pushed down through the corresponding branch. Now the instance reaches the leaf node, where it is classified as yes.

2.2 Experimentation

$$\textbf{RETICINDEX} = \textbf{RETICCOUNT} * \frac{\textbf{HEMATOCRIT}}{\textbf{NORMAL HEMATOCRIT}}$$

In a person whose reticulocyte count is 5%, hemoglobin 7.5 g/dL, hematocrit 25%, the RPI would be:

$$\text{RPI} = \frac{\text{ReticIndex}}{\text{Maturation Correction}}$$

$$\textbf{RPI} = \frac{\mathbf{5 * 25/ 45}}{\mathbf{2}}$$

RPI = 1.4

- The reticulocyte index (RI) should be between 1.0% and 2.0% for a healthy individual.
- RI < 2% with anemia indicates loss of red blood cells, but decreased production of reticulocytes (ie, and inadequate response to correct the anemia) and therefore red blood cells.
- RI > 3% with anemia indicates loss of red blood cells (from causes such as destruction, bleeding, etc.), with an increased compensatory production of reticulocytes to replace the lost red blood cells.

Microcytic anemia is primarily a result of hemoglobin synthesis failure/insufficiency, which could be caused by several etiologies. Macrocytic anemia, the most common cause of macrocytic anemia, is due to a deficiency of either vitamin B_{12}, folic acid, or both. Deficiency in folate and/or vitamin B_{12} can be due either to inadequate intake or insufficient absorption. Folate deficiency normally does not produce neurological symptoms. Normocytic anemia occurs when the overall hemoglobin levels are decreased, but the red blood cell size remains normal.

2.3 Attributes

1. Name of the person-(male/female)
2. Age (in yrs) - Meal (empty stomach, lunch, full)
3. Sex (M/F) -Amount of alcohol (ethanol in units)
4. Height and weight-(person height and weight)
5. Mass (in kg) -Blood alcohol content (high/low)
6. Tobacco use- Blood pressure (high/low)
7. Height (in cm) - Time duration (time spent drinking)
8. Complete blood cout (CBC)

Table 1: Attributes for measurement

2.5 EXAMPLES FOR ANEMIC MESUREMENT

Example 1: If_then_rule induced in the diagnosis

IF Sex = MALE
AND hemoglobin = 8.9
AND hematrocrit = 36
THEN
Diagnsis = Anemia.

Causes of Anemia

- Do not have adequate iron
- An impaired ability of the digestive tract to absorb the B12
- Blood loss are at risk of developing iron deficiency, such as women blood loss during menstruation, people with ongoing gastrointestinal blood loss,etc
- Medications cause anemic
- Some types of anemia are due to inherited or genetic defects

Example 2: If_then_rule induced in the diagnosis

IF Sex = FEMALE
AND hemoglobin = 12
AND hematrocrit = 33
THEN
Diagnsis = NORMAL

We have applied this technique here because of the ready availability of subjects with some knowledge of the domain that can provide feedback on the explanations. This can be used for decision making in healthcar

4. RESULTS AND DISCUSSION

Using this technique, the attribute weight, sex, meal, time duration and amount produced the best results. The availability of huge amounts of medical data leads to the need for powerful data analysis tools to extract useful knowledge. Researchers have long been concerned with applying statistical and data mining tools to improve data analysis on large data sets. Disease diagnosis is one of the applications where data mining tools are proving successful results.

Age or gender group	Hemoglobin(g/Dl)	Hematocrit (%)
Children(0,5-4)	<11.0	<33
Children (5-12)	<11.5	<35
Children(12-15)	<12.0	<36
Adult men	<13.0	<39
Non-pregnant women	<12.0	<36
Pregnant women	<11.0	<33

5. CONCLUSION

Using single data mining technique in the diagnosis of many health problems has been comprehensively investigated showing acceptable levels of accuracy. In this research work we examined the anemic issues with the help of rule induction and decision tree algorithms; in future we apply this datasets to machine learning tool to predict the final results accurately with the help of WEKA tool.

REFERENCES

1. Frawley and Piatetsky-Shapiro, 1996. Knowledge Discovery in Databases: An Overview. The AAAI/MIT Press, Menlo Park, C.A.

2. Glymour, C., D. Madigan, D. Pregidon and P. Smyth, 1996. Statistical inference and data mining. Communication of the ACM, pp: 35-41.

3. Jones, A.W., 1990. Physiological Aspects of Breath-Alcohol Measurements. Alcohol Drugs Driving, 6:1-25.

4. Han, J. and M. Kamber, 2001. Data Mining: Concepts and Techniques. San Francisco, Morgan Kauffmann Publishers.

5. Weinstein, J., K. Kohn and M. Grever *et al.*, 1992. Neural computing in cancer drug development: Predicting mechanism of action. Science, 258: 447- 451.

6. Stanfford, G.C., P.E. Kelley, J.E.P. Syka, W.E. Reynolds and J.F. Todd, 1984. Recent improvements in and analytical applications of advanced ion-trap technology. Intl. J. ass Spectrometry Ion Processes, 60: 85-98.

7. Schnorrenberg, F., C.S. Pattichis, C.N. Schizas, K. Kyriacou and M. Vassiliou, 1996. Computer aidded classification of breast cancer nuclei.

8. Simon, B.P. and C. Eswaran, 1997. An ECG classifier designed using modified decision based neural networks.

9. Romeo, M., F. Burden, M. Quinn, B. Wood and D. McNaughton, 1998. Infrared microspectroscopy and artificial neural networks in the diagnosis of cervical cancer.

10. Ball, G., S. Mian, F. Holding, Allibone Ro *et al.*, 2002. An integrated approach utilizing artificial neural networks and SELDI mass spectrometry for the classification of human

tumors and rapid identification of potential biomakers. Bioinformatics, 18: 395-404.

11. Aleynikov, S. and E. Micheli-Tzanakou, 1998. Classification of retinal damage by a neural network based system.

12. Domine, D., C. Guillon, J. Devillers, J. Lacroix and J.C. Dore, 1998. Non linear neural mapping analysis of the adverse effects of drugs.

13. Sharma, A. and R.J. Roy, 1997. Design of a recognition system to predict movement during anesthesia. IEEE Transactions. *J. Computer Sci., 2 (2): 194-200, 2006* 200

14. Brickley, M.R., J.P. Stepher and R.A. Armstrong, 1998. Neural networks: A new technique for development of support systems in dentistry.

15. E. A.M. and H. Fu. Privacy preserving distributed learning clustering of healthcare data using cryptography protocols. In Computer Software and Applications Conference Workshops (COMPSACW), 2010 IEEE 34th Annual, pages 140 {145, july 2010.

16. K. Harleen and S. K. Wasan. Empirical study on application of data mining techniques in health care. Journal of computer science 2, pages 194{200, 2006.

17. A. Li, S. Wang, H. Zheng, L. Ji, and J. Wu. A novel abnormal ecg beats detection method. In Computer and Automation Engineering (ICCAE), 2010 The 2nd International Conference on, volume 1, pages 47 {51, february 2010.

18. L. Nevine and M. Malek. Data mining for cancer management in egypt case study: Childhood acute lymphoblastic leukemia. 2005. [12] N. I. of Health. About clinical practice guidelines.

19. N. R. T. and P. Jian. Introduction to the special issue on data mining for health informatics. SIGKDD Explor. Newsl., 9:1{2, June 2007.

20. van Driel M. A., C. K., K. P. P., L. J. A., and B. H. G. A new web-based data mining tool for the identi_cation of candidate genes for human genetic disorders. Eur J Hum Genet, 11(1):57{63+, 2003.

21. Houston, Andrea L. and Chen, et. al.. Medical Data Mining on the Internet: Research on a Cancer Information System. Artificial Intelligence Review 1999; 13:437-466.