International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Impact Factor (2012): 3.358

Identification of Various Deficiencies Using Data Mining Techniques – A Survey

D. Thangamani[*], P. Sudha[#]

[*]Research Scholar of Computer Science, [#] Assistant Professor, Department of Computer Science
Sree Saraswathi Thyagaraja College
Pollachi – 642 107, Coimbatore, Tamil Nadu, India

Abstract: ***Globalization and modernization changing people dietary patterns and life styles, in particular the nutrition transition away from fruits and vegetables and greater consumption of more energy dense, nutrient- poor diets dependence on television, computers and mobile phones foe leisure time along with reduced level of physical activity . This leads nutrition deficiency of protein, carbohydrates, fats, minerals and vitamins. This nutrition deficiency is major factor for many global burdens of diseases. Health care professionals and decision makers from government using data mining to analyze deficiencies from several healthcare surveys and medical records to improve public health. This study provide malnutrition identification from children and elder people, recognition of anemia by iron deficiency, hair loss diagnosis by analyzing zinc, iron, vitamin deficiencies and also it uses mining techniques of decision tree, classification methods, artificial neural networks, support vector machines and statistical methods.***

Keywords: Malnutrition, Deficiencies, Data mining, Decision tree, Classification, Artificial neural network

1. Introduction

The objective of data mining is to identify valid novel, potentially useful, and understandable correlations and patterns in existing data. Finding useful patterns in data is known by different names (including data mining) in different communities (e.g., knowledge extraction, information discovery, information harvesting, data archeology, and data pattern processing) [1]. Today data mining is primarily used by companies such as banking, insurance, medicine, retailing, and health care commonly use data mining to reduce costs, enhance research, and increase sales. The medical community sometimes uses data mining to help predict the effectiveness of a procedure or medicine. The huge amounts of data generated by healthcare transactions are too complex and voluminous to be processed and analyzed by traditional methods. Data mining can improve decision-making by discovering patterns and trends in large amounts of complex data. Such analysis has become increasingly essential as financial pressures have heightened the need for healthcare organizations to make decisions based on the analysis of clinical and financial data. Data mining techniques are being increasingly implemented in healthcare sectors in order to improve work efficiency and enhance quality of decision making process.

The growing elderly population has a huge impact on the healthcare system, and malnutrition is a common health problems found frequently among the elderly [6,7]. The leading causes of malnutrition include several individual factors, such as poor oral health, loss of vision and hearing, dementia, impaired mobility, and pain. Malnutrition is both a cause and consequence of many geriatric diseases that accounts for a significant proportion of national medical spending [8]. Nutritional status has a great influence on the immune system of the human body; the lower immune response induced by malnutrition eventually puts elderly individuals at a high risk of infection, increasing the risk of mortality [9]. A rapid and precise diagnostic method that supports clinical decision enables the earlier identification of the elderly at risk of malnutrition. This will eventually prevent negative outcomes caused by poor nutritional status, resulting in a substantial reduction in healthcare cost spent on the elderly population [10]. A decision tree model presented in this study was developed from a combination of decision tree approaches and statistical analysis. Data mining provides information that can be the used in the analysis of risk factors for certain types of diseases.

Data mining is an area which used in vast field of areas. Rule based classification is one of sub areas in data mining. Rule based classification is used alone with Agent Technology to detect malnutrition in children. This proposed system is implemented as an E-government system. It will try to research whether there is connection between numbers of rules which is used with the optimality of the final decision [3].

Machine learning procedure offers a major platform in cases where a model lacks and the amount of data is enormous in explaining the relation and the generation of the data that is set. A research on trends and application of machine learning such as algorithms, techniques, and methods present practical functions for problem solving and application of techniques in settling and automatic data extraction. Anemia is one of the common diseases affecting individuals worldwide. In this study, due to lack of a distinct models, an artificial neural network (ANN) and support vector machines (SVM) have been put in place to establish a non linear function that is continuous and expresses the interdependency of the data collected and erythrocytes levels. This study will identify the use of artificial neural networks, support vector machines and statistical models and methods in the recognition of iron deficiency that leads to anemic conditions [4].

Hair lose diagnosis using artificial networks, it influences attributes of gender, age, gentic factors, surgery, pregnancy, zinc deficiency, iron deficiency and anemia [5].

2. Methods

Subjects

Myonghwa park et al [2] usess large data set from the 2008 national study of Korean elderly was conducted to provide accurate information regarding the characteristic of the elderly at risk of malnutrition in Korea. Xu Dezhi et al[3] uses Srilankan data base to provide e-government project to improve Srilanka peoples nutritional status by using rule based classification data mining techniques. Jameela Ali Akrimi et al [4] uses a survey on 2000 adult patients in an emergency department of complain of chest pains to detect anemia recognition. Ahmad Esfandiari et al [5] uses 384 individual records. These data are collected through interviews with the help of physicians by the attribute of age, gender, genetic factors, surgery, pregnancy, zinc deficiency, iron deficiency and anemia.

3. Data Mining Techniques

A. Decision Tree Models

Decision tree learning is a method commonly used in data mining.[1] The goal is to create a model that predicts the value of a target variable based on several input variables. An example is shown on the right. Each interior node responds to one of the input variables; there are edges to children for each of the possible values of that input variable. Each leaf represents a value of the target variable given the values of the input variables represented by the path from the root to the leaf.

There are many specific decision-tree algorithms. Such algorithms are (i) ID3 (Iterative Dichotomiser 3), (ii) C4.5 (successor of ID3), (iii) CART (Classification and Regression Tree), (iv) CHAID (CHi-squared Automatic Interaction Detector). Performs multi-level splits when computing classification trees[1]—(v) MARS: extends decision trees to better handle numerical data. (vi)Conditional Inference Trees. Statistics-based approach that uses non-parametric tests as splitting criteria, corrected for multiple testing to avoid over fitting. This approach results in unbiased predictor selection and does not require pruning.

B. Artificial Neural Network

An artificial neural network (ANN) is a computational model that attempts to account for the parallel nature of the human brain. An (ANN) is a network of highly interconnecting processing elements (neurons) operating in parallel. These elements are inspired by biological nervous systems. As in nature, the connections between elements largely determine the network function. A subgroup of processing element is called a layer in the network. The first layer is the input layer and the last layer is the output layer. Between the input and output layer, there may be additional layer(s) of units, called hidden layer(s). Fig.1 represents the typical neural network. You can train a neural network to perform a particular function by adjusting the values of the connections (weights) between elements.

All artificial neural networks are divided into two learning categories: supervised and unsupervised. In supervised learning, the network is trained by providing it with input and output patterns. During this phase, the neural network is able to adjust the connection weights to match its output with the actual output in an iterative process until a desirable result is reached. An ANN of the unsupervised learning type, such as the self-organizing map, the neural network is provided only with inputs, there are no known answers. The network must develop its own representation of the input stimuli by calculating the acceptable connection weights.

4. Results and Discussion

Myonghwa park et al[2] presented malnutrition identification used various decision tree algorithms like C5.0,C&R tree, QUEST,CHAID. Decision tree follows various steps to producing results: Trees utilized 48 attribute were identified from 52 input variable in these 7 variables were included chewing ability, level of life satisfaction, depression status, health status, number of disease diagnosis, difficulties in daily activities caused by pain, economic state and monthly income were identified to be most significant variables.

Rules were generated to identify the risk of malnutrition. The six decision rules were the following: 1) good and very good chewing ability, depressed, level of life satisfaction less than 2.89; 2) good and very good chewing ability, depressed, level of life satisfaction above 2.89, severe difficulty in daily activity caused by pain; 3) low chewing ability, living with spouse or children; low chewing ability, living with spouse or children, depressed; 4) low chewing ability, living with spouse or children, not depressed, number of disease above 2.5; 5) low chewing ability, living with spouse or children, not depressed, number of disease less than 2.5, subjective income status is poor.

The traditional statistical method and chi-square test were used to find out whether there are significant differences between the elderly at risk of malnutrition and normal subjects in the seven final variables identified by the C&R Tree algorithm. All seven variables showed statistical significance with a p-value of <0.05. Chewing ability, in particular, showed the biggest differences between well-nourished participants and malnourished participants. The majority of participants with low to moderate chewing ability had poor nutrition status (84.1% and 70.2%, respectively) while eight out of ten participants with either good or very good chewing ability were well-nourished. In addition higher proportions of individuals with depression, low life satisfaction, many difficulties in daily activities and poorly perceived economic status were found to be at risk of malnutrition.

Using the C&R Tree model, a well-designed prediction model was developed, which showed good performance in finding associated rules [11]. The C&R Tree uses a recursive partitioning method that provides a very simple representation that displays accumulated knowledge with a well-organized structure. To predict continuous dependent

variables (regression) and categorical predictor variables (classification), the C&R Tree builds a classification and regression tree [12]. Compared with other classification technique used for classification or regression of problems, there are many benefits that can be obtained using the C&R Tree [13]. The simplicity of the C&R Tree enables clinicians to make rapid classification of new clinical observations. Moreover, when there is a little prior knowledge, tree methods are known to be well suited for data mining tasks for data from healthcare settings [14].

The final results of the tree method demonstrated the usability of large public health data sets with good feasibility of decision model in the classification of elderly with malnutrition. Through repeated testing and refinement of data mining and the C&R Tree in particular, it is anticipated that new knowledge will be discovered by more sophisticated analysis of healthcare data at the community level. As a result of this study, a reliable decision support model was designed that provides accurate information regarding the characteristics of the elderly with malnutrition. The algorithm used to construct the decision tree showed high accuracy, and it is expected to facilitate the discovery of discriminatory knowledge for the targeted problem. The C&R Tree, which was based on the C&R Tree method, provided excellent discrimination of the characteristics associated with malnutrition in the elderly. This decision tree can be utilized to identify community residing elderly individuals who are at high risk of malnutrition; this will eventually contribute to significantly reducing healthcare costs spent on treating malnutrition and its complications in the elderly.

Xu Dezhi et al[3] Presented rule based classification to detect malnutrition in children. The technology of data mining is widely used in various fields. E-Government is a grand new domain in recent years. When E-Government system use to process data, it need to choose what data is useful and what kind of new information which can get from the log file or from the database [19]. As e-government is the use of Information Technology to free movement of information to overcome the physical bounds of traditional paper and physical based systems [17, 15]. Hence data mining techniques can be used to improve e-government systems. The current situation of e-government in Sri Lanka is still in the developing stage and there are minimum interactions between health sectors. E-government initiatives in Sri Lanka [18] by using Agent Technology and Data mining technique such as rule based classification.

The project called the medical information portal will provide static text based and multimedia based (video streaming) health education content in Sinhala and Tamil to citizens. From the remote medical consultation system (Vidusuwa) it handles the challengers faced by the patients in Sri Lanka with regard to inequality of resource distribution within the existing eHealth infrastructure [16].

In rule base classification technique rule extraction is one of the major sections. There are two broad classes of methods for extracting classification rules: 1) direct methods, which extract classification rules directly from data, and 2) indirect methods, which extract classification rules from other classification models, such as decision trees and neural networks.

In direct method there are several methods which can use to extract the rules from the data set. From the sequential covering algorithm rules are generated in a greedy fashion based on a certain evaluation measure. The algorithm extracts the rules one class at a time from the data set. The criterion for deciding which class should be generated first depends on factors such as class prevalence or cost of misclassifying records from a given class.

Multi-Agent System to Reduce Malnutrition Framework

Multi-Agent System to Reduce Malnutrition (MASRM) is developed under the influence of Java Agent Development Framework (JADE) runtime environment. Therefore each instance of the runtime environment, the “Container” has corresponding Agent Management System (AMS), Directory Facilitator (DF). AMS is responsible to provide unique names to each agent in the system and creation and termination of the agent. DF is the agent who provides a yellow pages service by means of which an agent can find other agents to achieve a specific service. Remote Monitoring Agent (RMA) is a system tool which represents all the actions in the platform in a Graphical User Interface (GUI). Therefore it consist options to control an agent’s life cycle, of an agent which is done by the AMS. The Message Transport System is also called the Agent Intercommunication Channel (AIC), is the software component controlling the exchange of all messages within the platform, including messages to and from remote platforms. It will make sure that all the messages have the Agent Communication Language (ACL) format which is issued by FIPA.

System Architecture

MASRM was developed on top of JADE, which is suitable to operate in a heterogeneous, networked environment such as the Internet to provide wide-area health detection and health information service to the citizen in the country. The framework further improved by introducing another agent called Resource Agent to the framework. Fig. 1the system consists of six types of architectural components which will help to identify malnutrition children in the real world: (1) User Identify Agent (2) Detection Agent (3) Advice Agent (4) Knowledge – based data server (5) Medical Agent (6) Resource Agent. It illustrates the structure of a container in the system. Another agent called Reporting Agent will only function in the main container of the system.

MASRM system which will be developed under the e-government initiative to provide advice regarding nutrition as well as to provide an easy way for citizens to check nutrition status of their child. From this research paper it has further improve the MASRM system by in cooperating rule based classification technique to detect malnutrition. Further from this research it is highlighted that there is an

effect on number of rules which is used to make the final decision with the optimality of the final decision.

Jameela Ali Akrimi et al [4] shown their research about Anemia is one of the common diseases affecting individual's worldwide. In this study, due to lack of a distinct models, an artificial neural network (ANN) and support vector machines (SVM) have been put in place to establish a non linear function that is continuous and expresses the interdependency of the data collected and erythrocytes and leucocytes levels, networks of neurons are built, taken through cross validation with the use of Excel 4.32 software for neuron-solution hence, forming the artificial neural network (Suzuki, 2011). The performance of neural networks and other learning methods has been established as an effective in the prediction, determination and classification of clinical outcomes involving blood complications. Consequently, it shown that it is more accurate as compared to linear regression. The only way to determine the levels of these blood components is looking at aspects such as GIB, blood losses, dialysis efficiency, vitamin B12 deficiency, iron status, folic acid deficiency, pro-inflammatory cytokine activities, aluminum toxicity, and any previous treatments with angiotensin. With the use of ICD-9 codes of GIB, all the available variables that are needed to test and develop various learning methods are identified. This data; demographic data, signs and symptoms, laboratory data, endoscopic diagnosis, co morbidities and outcomes were collected and utilized to create a relative analysis of these models. The methods set out a high ability to generalize symptoms that have come into a machine in the form of data use the polynomial input method of transformation. All methods gave specific but correct outcomes.

All the machine learning methods provided considered in monitoring the functional status and individual sense of well being which are generally referred to as the measurement of the quality of life (Kopple & Massry, 2004).In this study of accuracy and effectiveness of SVM, ANN and statistical models in the diagnosis of iron deficiency, the optimum conditions for a stable hemoglobin level has to be maintained in the range of between 11 to 12 g/dl as being the recommended level, and the concentration of the hemoglobin set above 12 g/dl. However, considering the possibility of thrombotic activities, it should not go above 14 g/dl (Quaglini Barahona, & Andreassen, 2001).

The ANNs accuracy and maintaining of this accuracy when data required is unavailable suggests the importance of artificial neural networks in being a potential aid in diagnosis of anemic conditions during patient evaluation (Baxt, Shofer, Sites & Hollander, 2002). An example of the range that can be used is input layers sum up to17 units, 15 units in the hidden layer, and 8 units for the output layer. The highest performed results were obtained when the hidden layer units were 15, 0.7 learning rate, and 0.1 momentum. Consequently, it emerges that there is 71.56 percent testing and 72.78 percent correctness. This shows that the potential of multilayered perception in the recognition and predicting of anemic cases and levels can be used by medical staff and hematologists (Suzuki, 2011).

The exclusion criteria used in the support vector machine include erythropoietin or therapy of androgen, presence of hepatic and inflammatory diseases, and blood transfusion in recent months. The major classification was either suffering or not suffering from iron deficiency (Quaglini Barahona, & Andreassen, 2001). This is in accordance with the response of administering iron therapy, leading us to a dichotomous response identified by NR for No response meaning no iron deficiency and R for Response to iron deficiency. The variables used included serum ferritin, red cells (GR), hemoglobin (Hb), mean corpuscular volume (MCV), iron binding capacity, serum iron, and hematocrit. The problem with this classification was faced with the theory of the SVMs involving linear approach, instead of ensuring that the errors are minimized in the training data.

Ahmad Esfandiari et al[5] used various Artificial neural networks algorithms to diagnosis the hair loss. Several algorithms have been introduced recently to train neural networks. These algorithms are mainly based on the standard method of error back-propagation the learning algorithms Levenberg-Marquardt, resilient back-propagation different algorithms of conjugate gradient such as Fletcher reeves, Polak-Ribiere, Rowel-Beale and scaled conjugate.

Hair loss has several types and each has different reasons. Reasons like genetic factors, diseases, poor nutrition, vitamin deficiencies, anemia, iron deficiency, stress, using cosmetics and so are effective on hair loss [25]. These factors are mentioned more fully in the next section. To recognize the amount of hair loss, artificial neural networks are used. Some effective factors in hair loss are used in neural networks as input parameters that the physicians and hair experts consider as effective. The actual data have been collected from several physicians who worked at clinics.

In this paper, they tried to determine the amount of hair loss through the significant characteristics of age, sex, genetic factors, pregnancy, surgery record, zinc deficiency, iron deficiency and use of cosmetics. The data for this study were gathered through interviews with doctors and conducting exact medical tests. The data were collected from 384 individuals and by getting help from physicians and specialists in one year. These factors are the input parameters of neural networks. We used the criteria Mean Square Error (MSE) and Mean Absolute Error (MAE) to evaluate the performance of the network.

The results obtained from comparison two-layer neural network, Levenberg- Marquardt has fewer prediction errors in comparison with other algorithms. The MSE of this algorithm in 19 epochs is equal to 0.0348 and its MAE is equal to 0.1393. Also, the MSE and MAE error of the Resilient back propagation algorithm in 30 epochs are equal to 0.0419 and 0.1584, respectively. Due to the fact that hair loss recognition rate is dependent on several factors and it is not easy to determine the exact cause, the results in this paper are satisfactory. The results from the algorithm were acceptable and among the different

algorithms the Levenberg-Marquardt algorithm had minor errors in fewer epochs and resulted in better outcomes.

References

[1] Margaret H. Dunham and S. Sridhar, Data Mining Introductory and Advanced Topics, ISBN: 0130888923, Pearson Education, Inc., Copyright 2003.
[2] Myonghwa Park, PhD, Hyeyoung Kim, PhD, Sun Kyung Kim, MSN," Knowledge Discovery in a Community Data Set: Malnutrition among the Elderly ", Healthcare Informatics and Research,2014.
[3] Xu Dezhi and Gamage Upeksha Ganegoda , "Rule Based Classification to Detect Malnut,rition in Children", International Journal on Computer Science and Engineering (IJCSE),2011
[4] Jameela Ali Akrimi, Abdul RahimAhmad, Loay E. George (IJSR), "Review of Machine Learning Techniques in Anemia Recognition", International Journal of Science and Research (IJSR), India Online ISSN: 2319-7064, 2013.
[5] Ahmad Esfandiari , Kimia Rezaei Kalantari and Abdorreza Babaei," Hair Loss Diagnosis Using Artificial Neural Networks", IJCSI International Journal of Computer Science Issues, Vol. 9, Issue 5, No 2, September 2012,ISSN (Online): 1694-0814.
[6] Studnicki J, Hevner AR, Berndt DJ, Luther SL. Comparing alternative methods for composing community peer groups: a data warehouse application. J Public Health Manag Pract 2001;7(6):87-95.
[7] Berndt DJ, Hevner AR, Studnicki J. Data warehouse dissemination strategies for community health assess-ments. Upgrade 2001; 2(1):48-54.
[8] Jang SN, Cho SI, Chang J, Boo K, Shin HG, Lee H, et al. Employment status and depressive symptoms in Koreans: results from a baseline survey of the Korean Longitudinal Study of Aging. J Gerontol B Psychol Sci Soc Sci 2009; 64(5):677-83.
[9] Kim HS. Income in old ages and role of children. Pro-ceedings of the 1st Korean Retirement and Income Study (KReIS) Conference; 2008 Jun 24; Seoul, Korea.
[10] Cropper S. Collaborative working and the issue of sustainability. In: Huxham C, editor. Creating collaborative advantage. London: SAGE Publications; 1996. P.80-100.
[11] Barrocas A, White JV, Gomez C, Smithwick L. Assessing health status in the elderly: the nutrition screening initiative. J Health Care Poor Underserved 1996;7(3):210-8.
[12] Garcia S, Fernandez A, Herrera F. Enhancing the ef-fectiveness and interpretability of decision tree and rule induction classifiers with evolutionary training set selection over imbalanced problems. Appl Soft Comput 2009; 9(4):1304-14.
[13] Huh MH, Lee YG. Data mining modeling and case. 2nd ed. Seoul: Hannarae; 2008.
[14] Koh HC, Leong SK. Data mining applications in the context of casemix. Ann Acad Med Singapore 2001; 30(4 Suppl):41-9.
[15] Babita, G., Subhasish, D., & Atual, G. (2008). "Adoption of ICT in a government organization in a development country: An empirical study." Stratefic Information Systems, 140-154. in press
[16] Dr. Keith R. P. Chapman MD, D. S. (2010). "ViduSuwa – Electronic Distant Healing: A Patient Centric Telemedicine Solution in SriLanka." Sri Lanka Journal of Bio-Medical Informatics 2010 , 63-75. in press
[17] Gupta, B., Dasgupta, S., & Gupta, A. (2008). "Adoption of ICT in a government organization in a developing country: An empirical study." Stategic Information Systems. in press
[18] Malnutrition in Sri Lanka. (n.d.). Retrieved May 25, 2009, from Unicef: http://www.unicef.org/srilanka/activities_1667.htm
[19] Yilei Wang, H. P. (2007). "The Data Mining of the E-Government on the Basis on Fuzzy Logic". International Conference on Integration Technology (pp. 774 - 777). Shenzhen, China: Proceedings of the 2007 IEEE.
[20] Kenji Suzuki. (2011). Artificial Neural Networks-Methodological Advances and Biomedical Applications. Introduction to the Artificial Neural Networks. Retrievedfrom<http://www.ltfe.org/wpcontent/uploads/2011/04/Artificial_Neural_Networks__Methodological_Advances_and_Biomedical_Applications.pdf>
[21] Bornn, L., & Tabet, A. (2010). Comment on "Particle Markov Chain Monte Carlo Methods". Journal of the Royal Statistical Society Series B, 72, 269-342.
[22] Quaglini, S., Barahona, P., & Andreassen, S. (2001).Artificial intelligence in medicine: 8th Conference on Artificial Intelligence in Medicine in Europe, Berlin: Springer.

Author Profile

D. Thangamani, M.Sc (CS & IT), Research Scholar, Sree Saraswathi Thyagaraja College, Pollachi , Tamilnadu, India

P. Sudha, MCA., MPhil., Assistant professor, Department of Computer Science, Sree Saraswathi Thyagaraja College, Pollachi – 642 107, Coimbatore, Tamil Nadu, India