Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X

IJCSMC, Vol. 4, Issue. 7, July 2015, pg.537 – 543

RESEARCH ARTICLE

USING DECISION TREE INDUCTION TO PREDICT THE TYPE OF EATING DISORDER

Ms. N.Vijayalakshmi[1], Ms. C.Vidhya[2]

[1]Asst. Professor, Dept. of M.C.A., Shrimati Indira Gandhi College, Trichy – 2, India

[2]Research Scholar in Computer Science, Shrimati Indira Gandhi College, Trichy-2, India

[1]E-mail: nvijimca@gmail.com ; [2]E-mail: vidhyachinna20@gmail.com

Abstract:* Decision Trees can be used to describe the features of a dataset. They can also be used to predict the class to which a specific data record belongs. In this paper, Information Gain applied to a patient database was used to induct decision trees. Three different kinds of decision trees were generated for the same data. The performance of each against the others was studied. The factors used for the study were grouped into three categories and the principal factors among each group were identified. Based on the determining factors under each category, a prediction model was created to predict the type of eating disorder to which the patient could be related based on his symptoms. The accuracy of the model was also measured. Frequent itemset mining and association rule mining were used to discover significant facts on patient tendencies over the period of study.

Keywords: Classification, Decision tree induction, prediction, frequent item set mining, association rule mining

I. INTRODUCTION

Data Mining is the new tool we use to mine information from empirical databases. In this research project, we have tried to discover certain facts about factors leading to anorexia nervosa, bulimia nervosa, and similar eating disorders. A sample population of 55 patients diagnosed with eating disorders and related symptoms has been taken. 217 records of these patients, 4 records per patient taken over 4 years, have been taken for study. Physiological factors, Psychological factors and socio-economic factors were taken and the effect of each of these factors on the population was also studied. General tendencies in transition between different disorders among the given cases have also been studied by frequent itemset mining and an attempt to carry out association rule mining is also done. Classification trees were inducted using J48, Random Tree and ID3 techniques for each of the group of factors and compared based on performance. Based on the results of the classification of the database using the training set, a prediction model was developed in C to diagnose the type of disorder on the test set and the results were compared with the actual diagnosis. An acceptable level of accuracy shows that the prediction model is good enough for use.

II. LITERATURE REVIEW

Study of various research activities related to eating disorders reveals that data mining has not been used for analysis of facts relating to this disease. In a study made by Vanderham et al.(1) criteria were mainly based on clinical impressions and on descriptive and inferential studies. Patients were grouped on an empirical basis, using principal components analysis (PCA) with optimal scoring (scaling), i.e. PCA with no a priori assumptions. Clustering was based on Morgan-Russell subscales, each measured four times over the course of illness. Contrary to DSM-IV criteria, patients did not cluster primarily on the basis of anorectic symptoms. The occurrence of bulimic symptoms was more dominant. Core symptomatology (preoccupation with food, disturbed body perception and inadequate sexual behaviour) did not differ between patients, either at referral or over time. These results support the spectrum hypothesis of the eating disorders, which considers them as one syndrome with different manifestations.

In another study made by Seidenfeld(2), it has been found that dieting behaviors and nutrition can have an enormous **impact on the gynecologic health** of adolescents. Teenaged patients with anorexia nervosa can have hypothalamic suppression and **amenorrhea.** Approximately one half of adolescents with bulimia nervosa also have hypothalamic dysfunction and oligomenorrhea or **irregular menses**.

The **Eating Attitudes Test** is one of the most widely used self-report eating disorder instrument. Because the EAT has not been validated with Diagnostic and Statistical Manual of Mental Disorders (4th ed. [DSM-IV]; American Psychiatric Association, 1994) criteria, its criterion validity for discriminating between nonclinical women with and without an undifferentiated DSM-IV eating disorder diagnosis(3) was examined. Differences in mean EAT scores among eating-disordered, symptomatic, and asymptomatic participants were also studied. Results show that the EAT has an **accuracy rate of at least 90%** when used to differentially diagnose those with and without eating disorders and that mean EAT scores differed among eating-disordered, symptomatic, and asymptomatic participants.

Clinical experience has suggested that women with eating disorders (ED) are prone to displace **negative feelings about the self** onto the body. This study (4) sought to evaluate these clinical observations by examining emotional inhibition and personality traits in women with ED. Female inpatients and intensive outpatients diagnosed with anorexia nervosa or bulimia nervosa were compared to women without an ED. The results of the study indicate that participants with ED inhibit their expression of both positive and negative emotions, even after controlling for neuroticism. Women with ED also reported higher levels of hostility and neuroticism. In addition, participants with ED were less aware of their inner thoughts and feelings (**private self-consciousness**) and had a heightened awareness of the thoughts and expectations of others (**public self-consciousness**). Finally, women with bulimia nervosa reported higher levels of emotional inhibition, neuroticism, public self-consciousness, and hostility when compared to women with anorexia nervosa. These data suggest that individuals who are not able to identify, and consequently, express their emotions may learn to handle emotional distress, interpersonal conflicts, and unexpressed hostility by turning their expression and lack of insight inward (viz., feeling “fat”).

Past research on eating disorders has concentrated on medical and psychological facets. Neglected in the literature are the social processes which antedate and maintain anorexia nervosa and bulimia. Women, the *primary* targets of eating disorders, respond to their **visual objectification by society** through **striving for thinness** and, in the extreme, through **starving or binging/purging**. Utilizing informal interviews as well as a two-year participant observation of an eating disorders self-help group, a study explores the meanings respondents attach to their eating anomalies(5). In their development of deviant identities, anorexics and bulimics proceed through the sequence of conforming behavior, primary deviance, and secondary deviance.

In another study(6), the authors present a method for studying Minuchin's **family interaction** concepts ‘enmeshment’, ‘rigidity’, ‘over protectiveness’ and ‘lack of conflict resolution’. The research procedure proposed consists of a series of standarized interaction tasks which are analyzed according to a behavioral coding system. The investigation method has been tried out in a pilot study of **ten families with an anorexia / bulimia nervosa patients**. The preliminary results appear to **support** the hypothesis that Minuchin's rather static family typology should be replaced by a more dimensional and **dynamic approach** of family functioning.

In yet another study(7), the development and validation of a new measure, the **Eating Disorder Inventory** (EDI) is described. The EDI is a 64 item, self-report, multiscale measure designed for the assessment of psychological and behavioral traits common in anorexia nervosa (AN) and bulimia. Reliability (internal consistency) is established for all subscales and several indices of validity are presented. First, AN patients (N = 113) are differentiated from female comparison (FQ subjects (N = 577) using a cross-validation procedure. Secondly, patient self report subscale scores agree with clinician ratings of subscale traits. Thirdly, clinically recovered AN patients score similarly to FCs on all subscales. Finally, convergent and discriminant validity are established for subscales. The EDI was also administered to groups of normal weight bulimic women, obese, and normal weight but formerly obese women, as well as a male comparison group. Group differences are reported and the potential utility of the EDI is discussed.

To investigate eating disorders (EDs) prevalence rates among Asian populations and identify characteristics that distinguish them from their Western counterparts, potential references were identified through an English-language literature search using Medline, Psychinfo, Dissertation Abstracts (1966 to 1999) and through extensive manual searching of textbooks, reviews and reference lists.(8). It was concluded **that EDs in Asian populations have received little attention** because they have been **predominantly viewed as associated with Western culture**. Classified by many as a "culture bound syndrome" of the West, they may really be a culture-change syndrome.

A study by Crisp H. Arthur(9) reviews risk factors for anorexia, bulimia, and obesity and discusses the use of educational, behavioral, and personal/experiential prevention efforts. The case illustrates the origins of massive obesity and **changes in mood, activity, and family functioning** concomitant with treatment. Risk factors identified include childhood weight problems, enmeshed family systems, and possible genetic influences. Secondary prevention is considered in relation to immediate motivating factors (morbid concerns about weight) and predisposing maturational problems. Primary prevention is discussed in terms of developmental determinants of obesity, binge eating and avoidance behavior, dysmorphophobia, and personal development. (PsycINFO Database Record (c) 2012 APA, all rights reserved)

III. CASE ANALYSIS

A group of 55 people suffering from eating disorders were taken for study. The patients were diagnosed once every year for a period of 4 years. The data pertaining to this control group was recorded during every visit. 3 records are missing as data relating to 3 controls could not be recorded as they failed to visit for diagnosis during one of the years. Therefore totally 217 records are available for study.

16 different parameters are considered for every person during every visit. The parameters can be categorized into 4 groups. Eating habits like fasting, vomiting after eating, binge-eating, purging after food form one group. Social behavior like family relationships, dependence of family, relationships in public places like at work/school, friendship form one group. Psychological behavior like moody behavior, sexual attitude, sexual behavior, hyperactivity, preoccupation with food and weight and body perception form one group. Independent factors like weight and menstruation cycle form one group.

The type of eating disorder diagnosed is recorded for every patient at every visit at a 4 point scale.

1	2	3	4
anorexia	anorexia with bulimia	bulimia	eating disorder

IV. METHODOLOGY

A. Frequent item set mining

Frequent itemset mining leads to the discovery of associations and correlations among items in large transactional or relational data sets. This can help in many decision-making processes. A group of elements that represents together a single entity is called an itemset. "A frequent itemset is one that occurs in at least a user-specified percentage of the database."

In data mining, association rule learning is a popular and well researched method for discovering interesting relations between variables in large databases. The problem of association rule mining is defined as: Let $I = \{i_1, i_2, \ldots\ldots i_n\}$ be a set of n binary attributes called *items*. Let $D = \{t_1, t_2, \ldots.t_m\}$ be a set of transactions called the *database*. Each transaction in D has a unique transaction ID and contains a subset of the items in. A *rule* is defined as an implication of the form $X \Rightarrow Y$ where $X, Y \subseteq I$ and $X \cap Y = \emptyset$. The sets of items (for short *itemsets*) X and Y are called *antecedent* (left-hand-side or LHS) and *consequent* (right-hand-side or RHS) of the rule respectively.

To select interesting rules from the set of all possible rules, constraints on various measures of significance and interest can be used. The best-known constraints are minimum thresholds on support and confidence. The *support* $supp(X)$ of an itemset X is defined as the proportion of transactions in the data set which contain the itemset. The *confidence* of a rule is defined as $conf(X \Longrightarrow Y) = supp(X \cup Y)/supp(X)$.

B. Association Rule Mining

Association rules are usually required to satisfy a user-specified minimum support and a user-specified minimum confidence at the same time. Association rule generation is usually split up into two separate steps:

1. First, minimum support is applied to find all *frequent itemsets* in a database.
2. Second, these frequent itemsets and the minimum confidence constraint are used to form rules.

C. Classification technique

A classification problem consists of properly segregating records into specific classes based on the attribute values. To build a classification model, the labeled data set is initially partitioned into two disjoint sets, known as the training set and the test set respectively. The training set consists of a relatively larger number of records that aid in creating a proper model for classifying the data based on the values of various attributes of the record. Accuracy of classification is one of the important parameters for classification. The test set is used to evaluate the accuracy of classification. A classification model can be used as an explanatory tool for distinguishing objects of different classes. Then it is called as a descriptive model. The model can also be used to predict the class labels of new records. Then it is called as a predictive model. There are a number of classification techniques available like decision trees, rule-based methods, neural networks, genetic algorithm and naïve Bayesian Belief Networks.

D. Decision Trees

A decision tree is a hierarchical structure of nodes and directed edges. The root node represents the entire unclassified data set. Internal nodes represent decisions involving values of attributes in the record set and branches represent an outcome of the test. Based on the values of specified attribute in the decisions, we take one directed outgoing path from each internal node. Leaf nodes represent classes and are labeled with the name of the class. There are many specific decision-tree algorithms. Notable ones include ID3, J48, CHAID, Random Trees etc. Decision trees are used in data mining in two main ways:

- To predict the class to which an entity record belongs based on its' attribute values. This is called as classification tree analysis.

- To predict the value of an attribute based on other attribute values of the entity record. This is called as regression tree analysis.

E. Decision Tree Induction

Algorithms for constructing decision trees usually work top-down, by choosing a variable at each step that best splits the set of items. The basic strategy of the algorithm is to determine how to split the records at every step. We could use some greedy heuristics to make a series of locally optimum decisions about which attribute to use for partitioning the data. Used by the ID3, C4.5 and C5.0 tree-generation algorithms, Information gain is based on the concept of entropy from information theory. Information Gain is defined as the difference between the original information requirement based on the proportion of the classes and the new information requirement obtained after partitioning on A. The attribute giving the maximum Information gain is chosen for splitting, after checking with all possible attributes that have not been used for splitting.

```
Assume there are $k$ classes $C_1$, ... $C_k$ ($k$ = 2 in our example).
// to decide which attribute to split on:
    • for each attribute that has not already been used
        o Calculate the information gain that results from splitting on that attribute ie., Info-Gain$_A$(T)
        o Split on the attribute that gives the greatest information gain out of all the attributes
```

V. RESULTS AND DISCUSSION

Data mining was applied to this data set to find the following facts:

During the first year 36 persons suffered from anorexia, 5 from anorexia with bulimia, 11 with bulimia and 3 with atypical eating disorder. During the next 3 years there was a transition from each of the types to other types for a subset of patients. It was observed that the no. of patients suffering from anorexia lessened and those with eating disorder increased significantly over the period of four years.

The year wise change can be viewed better in the table I.

TABLE 1

NUMBER OF PATIENTS DIAGNOSED WITH EACH TYPE OF EATING DISORDER FOR EACH YEAR

	ANOREXIA	ANOREXIA WITH BULIMIA	BULIMIA	EATING DISORDER
YEAR1	36	5	11	3
YEAR2	20	6	9	20
YEAR3	19	6	4	26
YEAR4	16	1	9	29

The change in type of eating disorder during every year was also studied for every pair of eating disorders. We can see that the value for 11 that is no change from anorexia nervosa and 44 that is no change from typical eating disorder, score the maximum among transitions in the table. This implies that the greatest probability of change from one type to another is taken by 11 and 44. Another observation from the given data is that 2% of the 55 cases have been out of risk ie, at type 4 eating disorder during all the four years. 23% have been improved after 1 year, 13% after 2 years and 17% after 3 years. Therefore totally 29 persons, that is 56% of the total no. of controls have been improved in their status through the four years. This shows that the treatment given in terms of clinical tests, medicines and psychiatric counseling have borne fruit through the period.

A. Frequent itemset mining from the dataset:

Applying frequent itemset mining on the data set shows that no. of times anorexia was diagnosed is 91 out of 220 that is 41%. And no. of times atypical eating disorder was diagnosed is 78 out of 220 ie 35%. Assuming a support value of 30% we move to 2 itemset mining. The only possible two itemsets are 14 and 41. However the no. of occurrences of these two itemsets is very low as shown in the table. Assuming a confidence of 15% we see that the case 14 satisfies our requirement. 17% of the time anorexia leads to atypical eating disorder. **Therefore we mine the rule that a person with anorexia has a better chance of moving to eating disorder, a case out of risk.**

B. Application of classification techniques on the given dataset:

The 16 parameters have been grouped under 3 heads. Fasting, Binge-eating, Vomiting and Purging were taken together as eating habits group. Next relationship with family, emancipation from family a, regularity at school / workplace, and number of friends were grouped together to form a social-relationship group. Sexual attitude, Sexual behavior, mood, pre-occupation with body and food, hyperactivity and feeling on their body appearance were bundled to form psychological factors group. The weight and menstrual period information was treated separately**.**

Each group was taken together with the already diagnosed type of eating disorder and Iterative Dichotomizer 3 Algorithm was applied on each group manually using information gain method to identify the attribute that best classified the group at every stage. Using this a decision tree was generated for each of the three groups. Weka 6.2 data mining tool was used for applying classification methods on the same groups. J48 classification tree and Random Tree methods were used on the same groups to classify the data records and generate decision trees. Table II shows the classification accuracy obtained by each technique.

TABLE II

CLASSIFICATION ACCURACY OBTAINED THROUGH EACH METHOD

	J48	Random Forest	ID3
Classification Accuracy	49%	47%	67%

Using the trees generated by the ID3 method, we have generated an application in C that would predict the eating disorder, based on the inputs given for the 16 factors considered. The prediction accuracy of the application has also been found to be 67%.

VI. CONCLUSION

ID3 was found to be better than J48 and Random Forest techniques in classifying the given dataset. J48 produced short and well pruned trees with minimum number of nodes. Random Forest produced overfitted trees with large number of levels and nodes. ID3 produced a well balanced tree with pruning done at lower levels. The classification accuracy of the three types of techniques shows that ID3 is the best of all. The prediction model could correctly predict the type of eating disorder upto an accuracy of 67%. This shows that the classification tree inducted has well captured the features of the dataset for diagnosis.

REFERENCES

1. Van der Ham, T., J. J. Meulman, D. C. Van Strien, and H. Van Engeland. “Empirically based subgrouping of eating disorders in adolescents: A longitudinal perspective”, *British Journal of Psychiatry,* Issue 170, 363-368, 1997.

2. Seidenfeld ME, Rickert VI,” Impact of anorexia, bulimia and obesity on the gynecologic health of adolescents”, *American Family Physician*, Vol.64(3),445-450, 2001

3. Laurie B. Mintz & M. Sean O'Halloran, “The Eating Attitudes Test: Validation With DSM-IV Eating Disorder Criteria”, *Journal of Personality Assessment*, Volume 74(3), 2000

4. Kelsie Forbush, and David Watson, Department of Psychology, The University of Iowa, Iowa City, IA, USA,” Emotional Inhibition and Personality Traits: A Comparison of Women with Anorexia, Bulimia, and Normal Controls”, *Annals of Clinical Psychiatry*, Vol. 18(2) , 115-121,2006

5. Penelope A. McLorg & Diane E. Taub ' “Anorexia nervosa and bulimia: The development of deviant identities”**,** *Deviant Behavior*, Volume 8(2), 177-189, 18 May 2010

6. Kog, E., Vandereycken, W. and Vertommen, H., “ Towards a verification of the psychosomatic family model: A pilot study of ten families with an anorexia/bulimia nervosa patient”, *International Journal of Eating Disorders,* Volume 4: 525–538, 1985.

7. David M. Garner, Ph.D., Marion P. Olmstead, M.A., Janet Polivy, Ph.D. “Development and validation of a multidimensional eating disorder inventory for anorexia nervosa and bulimia”,Vol 2(2), 15-34,1983

8. Grace Tsai, Ph. D., "Eating disorders in the Far East",*Eating and Weight Disorders –Studies on Anorexia, Bulimia and Obesity,* Voll. 5(4), 183-197, Dec. 2000

9. Crisp, Arthur H., "Some possible approaches to prevention of eating and body weight/shape disorders, with particular reference to anorexia nervosa", *International Journal of Eating Disorders,* Vol 7(1), 1-17, Jan 1988.