An Information Gain based Fuzzy Classifier for Predictive Analysis in Colon Cancer Data

N.S.Nithya
Dept of Computer Science and Engg
K.S.R.College of Engineering,
Tiruchengode, India,

Dr.K.Duraiswamy
Dept of Computer Science and Engg
K.S.R.College of Technology,
Tiruchengode, India

ABSTRACT
Modern medicine generates a great deal of information stored in the medical database. Extracting useful knowledge and providing scientific decision making for the diagnosis and treatment of disease from the database increasingly becomes necessary. In India most of the people suffering cancer diseases. Using association rule mining for constructing classification system for diagnosing cancer diseases is a promising approach. A detailed survey shows that a combined approach which integrates the Fuzzy weighted association mining and information gain method may be used to find the associated attribute based on information gain which assigns a weight value to support ,confidence measure and also a fuzzy association mining rule may be used to classify the cancer diseases. This approach would provide a better accuracy compared to other association rule mining technique.

Keywords
Information gain, weighted support and confidence, Fuzzy association rule mining.

1. INTRODUCTION
The diagnosis of diseases is a vital and intricate job in medicine. Colon cancer is a malignant cancer in which the tissue in the walls of colon/rectum develops abnormally. It is the third common type of cancer in both sexes. This cancer forms in the tissues of the colon (the longest part of large intestinal).Colon cancer is usually caused due to a diet high in fat, red meat but low in fruits and vegetables, high caloric intake, low levels of physical activity, obesity, smoking and excessive alcohol intake etc. Prediction of this colon cancer will help to prevent it in its early stage. Data mining techniques have been usually applied in this field (or) other aspects of medical science.

A classification system can assist the physician in this process. The system can predict if the patient [1] is likely to have a certain disease (or) present incompatibility with some treatments. Considering the output of the classification model, as in [2]. The physician can make a better decision on the treatment to be applied to this patient. Association rule mining find the correlated attribute for constructing classification system to predict this colon cancer. Given the readability of the associative classifiers, they are especially fit to applications were the model may assist domain experts in their decisions. Medical field is a good example for such applications. There are many associative classification approaches that have been proposed recently such as FWFP growth, FWARM, Utility based association mining, Quantitative approach and Constraint based association and Fuzzy partitioning ARM.

Fuzzy association rule mining plays an important role in medical data mining. The idea of empowering classical association rules by combining them with fuzzy set theory has already been around since several years. A problem of classical association rules is that not every kind of data can be used for mining. To overcome this problem, the approaches of fuzzy association rules have been developed. It allows the intervals to overlap, making the set fuzzy instead of crisp. Using this approach rules can be discovered that might have got lost with the standard quantitative approach.

2. INFORMATION GAIN MEASURE
Information gain measure as in [3]. is used for attribute selection. Based on the information gain assign the weight to different attribute and can get more accuracy in predictive modeling system like medical field etc. In any prediction model all attributes do not have same importance in predicting the class label. So different weights can be assigned to different attributes according to their information gain measure. Attribute with highest information gain as in [4] .chosen as the test attribute and assign highest weight value. Information theoretic approach minimizes the number of tests needed to classify an object. The entropy H(D) of a dataset D is a measure of the disorder/variation/information in it. If all the records in the dataset belong to the same class, then the entropy would be zero. If all the records are uniformly distributed among the different classes, the entropy would be maximized. The entropy H(D) of a dataset D whose records are divided into m classes with probabilities p1,p2…..pm is defined as

$$H(D) = -\sum_{i=1}^{m} p_i \log\ p_i \qquad (1)$$

The best split for each attribute is chosen based on a criterion known as information gain. Given that a dataset D is split into D_1, D_2, D_3……Dn the information gain of the split is computed as

$$Gain = H(D) \quad - \quad \sum_{i=1}^{n} P(D_i)H(D_i) \qquad (2)$$

In this equation, the first part is the entropy of the dataset before the split, whereas the second part is the average (or) expected entropy of the collection of the datasets after the split. The ID3

algorithm selects the split with maximum information gain. The medical dataset mostly contain categorical attributes. To make this approach feasible, the ID3 algorithm only considers categorical attributes because the number of distinct values is small and hence, can be enumerated. The Table-1 gives the causes of colon cancer according to information gain approach relevant attributes (potential causes of cancer) are selected. The ranking are given to the selected attributes in table-2 based on Information gain.

Table 1. Colon Cancer Dataset

S.No.	Attribute	Value
1	Age	Young, Middle ,Old
2	Sex	Male , Female
3	Smoking	Yes ,No
4	Alcohol Intake	High, Low
5	Diabetes	Yes, No
6	Family history of cancer	Yes, No
7	Obesity	Yes, No
8	Exercise	Yes, No
9	High red meat and fat Diet	Yes, No
10	Previous History of Colon Polyps	Yes, No
11	Previous History of Inflammatory bowel disease	Yes, No
12	Previous history of Cancer	Yes, No
13	Radiation Exposure	Yes, No
14	Life Style(Race and Ethnicity)	Yes, No
15	Anemia	Yes, No

Table 2. Selected Attribute

S.No.	Attribute
1	Age
2	Sex
3	Smoking
4	Alcohol Intake
5	Diabetes
6	Family history of cancer
7	Obesity
8	Exercise

3. BASED ON UW SCORE (Weightage and utility)

Utility based data mining as in [5]. is a research area entranced in all types of utility factors in data mining processes and focused at integrating utility considerations in data mining tasks. Utility of an attribute is a subjective term dependent on users and applications. It could be measured in terms of risk and potential cause of death related to medical field. The incorporation, weighted utility association rule mining (WUARM) can be considered as the extension of weighted association rule mining in the sense that it considers items weights as their significance in the dataset and also deals with the frequency of occurrences of items (symptoms) in the transactions.

3.1 Computation of w-gain

Item weight (Wi): Item weight is the quantitative measure of the attribute contained in the transaction database D. Item weight value Wi is a non negative integer.

Weighted gain (W-gain): w-gain is defined as the sum of item weights Wi of an attribute contained in every transaction of the database D.

$$W-gain \quad = \quad \sum_{i=1}^{|T|} W_i \tag{3}$$

where W_i is the item weight of an attribute and $|T|$ is the number of transactions in the database D.

3.2 Computation of u-gain

Item Utility(U_i): The item utility is generally defined as the margins of profit associated with that particular attribute. It is denoted by U_i.

Utility factor (U-factor): The utility factor is a constant that is determined by the sum of the all items utility **(U_i)** It is defined as

$$Utility \;\; factor = \frac{1}{\sum_{i=1}^{m} U_i} \tag{4}$$

Utility gain (U-gain): Utility gain refers to the measure of attribute's actual utility based on the U-factor.

$$U\cdots gain \quad = U_i \times \quad U\cdots factor \tag{5}$$

3.3 Computation of UWscore from W-gain and U-gain

$$\frac{\sum_{i=1}^{|R|} (w-gain)_i \; * \; (U-gain)}{|R|} \tag{6}$$

UW score is defined as the ratio between the sum of products of w-gain and U-gain for every attribute in the association rule to the number of attributes present in the rule. Where $|R|$ represents the number of attributes in the association rule.

4. FWFP GROWTH APPROACH (FUZZY WEIGHTED FREQUENT PATTERN GROWTH)

In this approach[6] combine the concept of fuzzy weight, fuzzy partition methods in data mining, and use FP growth to propose FWFP growth algorithm mining all association rules. Numbers of calculations are involved to calculate fuzzy average weight, fuzzy weighted support and confidence, and rebuild fuzzy set table. This method also gives better accuracy when it is applied to Apriori algorithm which reduces repeated scanning process for association rule mining. The triangle membership function constructs usefulness and validation in Fuzzy system. FWFP growth algorithm solves the disadvantages of Apriori algorithm in repeated scanning database and the waste of reducing time.

5. WEIGHTED ASSOCIATION RULE MINING

In this weighted associated rule mining as in [7].generalize the problem of downward closure property. The problem of invalidation of DCP is solved using an improved model of weighted support and confidence framework for classical and Fuzzy association rule mining. This method follows an Apriori algorithm approach and avoids pre and post processing as opposed to most weighted ARM algorithms, thus eliminating the extra steps during rule generation.Table-3 gives the sample dataset of colon cancer using this dataset calculate the Weighted support and confidence value.

Table 3. Sample Colon Cancer Dataset

S.No.	Age	Sex	Smoking habit	Alcohol intake	Diabetes
1	86	M	Yes	High	No
2	85	F	Yes	No	No
3	61	M	Yes	High	Yes
4	65	F	No	No	No
5	54	F	No	Low	No
6	85	F	No	No	No

5.1 Item Weight

IW is a value attached with each item set. It is a non-negative real number value range [0…1] with respective to some degree of importance.

5.2 Weighted Support

Weighted support as in [8]. WSP of X→classlabel, Where X is set of non empty subsets of attribute value set, is fraction of weight of the record that contains above attribute value set relative to the weight of all transactions.

$$WSP(X \rightarrow classlabel) = \frac{\sum_{i=1}^{|x|} weight\ (r_i)}{\sum_{k=1}^{|n|} weight\ (r_k)} \quad (7)$$

Example: Consider a rule R (Smoking=”yes”) →colon cancer =”yes” then weighted support of R is calculated as:

$$WSP(R) = \frac{(Sum\ of\ record\ weight\ having\ the\ condition\ smoking = "yes"\ true\ and\ also\ given\ class\ label\ colon\ cancer)}{sum\ of\ weight\ of\ all\ transactions} \quad (8)$$

5.3 Weighted Confidence

Weighted confidence as in[9]. of a rule X→Y where Y represents the class label can be defined as the ratio of weighted support of (X Y) and the weighted support of(X).

$$Weighted\ confidence = \frac{Weighted\ support\ \ (X \cup Y)}{Weighted\ Support\ (X)} \quad (9)$$

$$WC(R) = \frac{(Sum\ of\ record\ weight\ having\ the\ condition\ smoking = "yes"\ true\ and\ also\ given\ class\ label\ colon\ cancer)}{Sum\ of\ record\ weight\ having\ the\ condition\ smoking = "yes"\ true} \quad (10)$$

6. FUZZY ASSOCIATIVE CLASSIFIER

A fuzzy classifier [10] usually includes many fuzzy rules which state the relationship between the attributes in the antecedent part of a rule of the class label in the consequent part. The fuzzy classifier is a combination of fuzzy set and fuzzy clustering. In the antecedent part, there are many attributes which are mapped to the fuzzy sets by some continuous membership functions, and in the consequent part of rules, there is a crisp set of class labels.

If (n_1>50)Λ … Λ(n_m, yes), then class1. Where, n_1, $n_{2....}$ n_m are the attributes of the colon cancer dataset, Λ is an operator and 50, yes are quantitative attribute. So making a fuzzy classifier needs to convert quantitative data as in [11] into fuzzy membership functions and rule mining. To make the fuzzy membership functions, the quantitative attributes are categorized by clustering process. Clustering can be used as preprocessing step for mining association rules.

For example, When attribute R= {age, smoking}, a fuzzy item set can be like <age: middle U< smoking: high >Age is converted into fuzzy set as follows

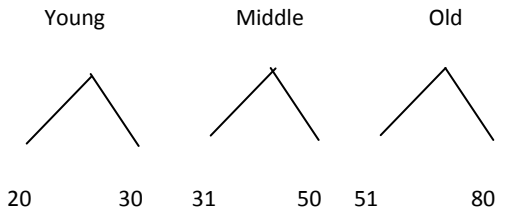

Fig 1: Fuzzy classification

After applying fuzzy sets, colon cancer data set is converted into fuzzy data set. After finding the clusters, apply the Apriori in each cluster in mining association rules. Apriori algorithm as in [12].is a well known method for rule mining which can be used in fuzzy association rule mining (FARM). In this algorithm, the attributes which are characterized with a common linguistic term and have a support value more than a predefined threshold are mined and make the antecedent of the rules. Fuzzy weighted support[13].is a well- known evaluation measure associated rule mining which shows the usefulness of a fuzzy attribute set. If FWS of a fuzzy attribute set is larger than a predefined value (λ), it is called a frequent fuzzy attribute set.

For example from table-3, If fuzzy 2 attribute set (X:A)= [Age: middle>u<smoking: High>]

For three samples be: (from table-3)
Sample1={(Age:middle/0.4), (smoking: High/0.8)}

Sample2={(Age:middle/0.5),(smoking:High/0.7)}

Sample3={(Age:middle/0.3),(smoking: High/0.9)} (11)

FWS (X:A) = 0.4 x 0.8 x weighted value + 0.7 x 0.5 x weighted value + 0.3 x 0.9 x weighted value.

Fuzzy weighted support is first used to find frequent item sets exploring its downward closure property to prune the search space. Then Fuzzy weighted confidence is used in a second step to produce rules from the frequent item sets that exceeds a min-confidence threshold. A Fuzzy association rule has a form like(X: A) (Y:B) where there is no interaction between (X:A) and (Y:B) is the consequent of a rule is a class label, the rule is considered as a classifier rule.

7. CONCLUSION

We have discussed various advanced Associative Classifiers being proposed in recent years. In future we combine a information gain approach based Fuzzy clustering associative classifiers. It gives better accuracy by comparing other associative classifiers. In the feature selection process, potential markers (potential causes of colon cancer) are selected based upon information gain and also assign a weightage value to support and confidence measure based on information gain for accurate classification. The non-additive of the Fuzzy measure reflects the importance of the feature attributes as well as their interactions. These two classifiers give explicit information on the importance of the individual mutated sites and their interactions toward the classification. Number of rules are produced by Associative classifiers are both relevant and irrelevant. The irrelevant rules are eliminated by the combination of information gain and Fuzzy clustering approach.

8. REFERENCES

[1] Jyoti Soni,Ujma Ansari, Dipesh Sharma and Sunita Soni. 2011. Predictive Data Mining for Medical Diagnosis:An Overview of Heart Disease Prediction, J. Comp. Appl.17(8).

[2] Ephziabah, E.P. 2011. Cost Effective Approch on Feature Selection Using Genetic Algorithms and Fuzzy Logic For Diabetes Diagnosis, J. Soft Comp. 2(1), pp. 1-10.

[3] Vikram Pudi and Radha Krishna, P. 2009. Data Mining, Oxford University Press.

[4] Ashraf, M., Kim Le and Xu Huang. Information Gain and Adaptive Neuro-Fuzzy Inference System for Breast Cancer Diagnoses,pp.911-915.

[5] Parvinder, S., Sandhu, Dalvinder,S., Dhaliwai and Panda, S.N. 2011. Mining Utility-Oriented Association Rules: An Efficient Approach Based on Profit and Quantity, J.Phy. Scie., 6(2), pp. 301-307.

[6] Chien-Hua Wang and Chin-Tzong Pang.2009.Finding Fuzzy Association Rule Using FWFP-Growth with Linguistic Supports and Confidences, J. Info.and Mathe. Sci., 5(4), pp. 300-308.

[7] Murtagh, Tao.F., Farid, F. 2003.M:Weighted Association Rule Mining Using Weighted support and siginificance Framework, Proceedings of 9th ACM SIGKDD Conference on Knowledge Discovery and Data Mining,pp.661-666.

[8] Ruijuan Hu. 2010. Medical Data Mining Based on Association Rules, J.Comp. and Info. Sci., 3(4).

[9] Sunitha Soni, Vyas, O.P. 2010. Using Associative Classifiers for Predictive Analysis in Health Care Data Mining, J. Comp. Appl.,4(5), pp.33-37.

[10] Saket Agarwal and Leena Singh.2011.Mining Fuzzy Association Rule Using Fuzzy Artmap for Clustering, Journal of Engineering Research and Studies, 2, pp.76-80.

[11] Vijay Krishna,V and Radha Krishna, P. 2008. A novel approach for statistical and fuzzy association rule mining on quantitative data, J. Sci. & Indu. Research, 67,pp. 512-517.

[12] Essam AI-Daoud.2010.Cancer Diagnosis Using Modified Fuzzy Network, Universal J. Comp. Sci. and Engg. Tech.,1(2),pp. 73-78.

[13] Maybin Muyeba, M.Sulaiman Khan and Frans Coenen. 2010. Fuzzy Weighted Association Rule Mining with Weighted Support and Confidence Framework, Scientific literature Digital Library and Search engine.