# Using Associative Classifiers for Predictive Analysis in Health Care Data Mining

Sunita Soni
Associate Professor
Bhilai Institute of Technology,
Durg-491 001, Chhattisgarh, India

O.P.Vyas
Professor
Indian Institute of Information
Technology, Allahabad- 211012 (U.P.), India

## ABSTRACT
Association rule mining is one of the most important and well researched techniques of data mining for descriptive task, initially used for market basket analysis. It finds all the rules existing in the transactional database that satisfy some minimum support and minimum confidence constraints. Classification using Association rule mining is another major Predictive analysis technique that aims to discover a small set of rule in the database that forms an accurate classifier. In this paper, we introduce the combined approach that integrates association rule mining and classification rule mining called Associative Classification (AC). This is new classification approach. The integration is done by focusing on mining a special subset of association rules called classification association rule (CAR). And then classification is being performed using these CAR. Using association rule mining for constructing classification systems is a promising approach. 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 was such applications may appear. Consider an example were a physician has to examine a patient. There is a considerable amount of information associated with the patient (e.g. personal data, medical tests, etc.). A classification system can assist the physician in this process. The system can predict if the patient is likely to have a certain disease or present incompatibility with some treatments. Considering the output of the classification model, the physician can make a better decision on the treatment to be applied to this patient. There are many associative classification approaches that have been proposed recently such as CBA, CMAR, CPAR and MCAR and MMAC. Also Combining the Advanced association rule mining with classifiers gives a new type of Associative classifiers with small refinement in the definition of support and confidence that satisfies the validation of downward closure property. We will discuss advanced associative classifiers being proposed in recent years to provide better accuracy as compare to traditional Classifiers.

## Keywords

Associative Classifiers, CBA, CMAR, CPAR, MCAR

## 1. INTRODUCTION
Data mining is a process, which involves the application of specific algorithms for extracting patterns (models) from data. New knowledge may be obtained in the process while eliminating one of the largest costs, *viz.*, data collection. Medical data, for example, often exists in vast quantities in an unstructured format. A new predictive modeling approach known as associative classification, integrating association Mining and classification inside into single system is being discussed as better alternative for predictive analytics [3]. Some of the classification techniques presented are CBA[10], CMAR[9], CPAR[8]. As discussed in [10] it achieves higher classification accuracy than do traditional classification approaches such as C4.5, FOIL, RIPPER. According to [10] these traditional classifiers are faster but in many cases accuracy is not so high. Moreover many of the rules found by associative classification method cannot be discovered by traditional classification algorithm. 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 were such applications may appear. Let us consider an example were a physician has to examine a patient. There is a considerable amount of information associated with the patient (e.g. personal data, medical tests, etc.). A classification system can assist the physician in this process. The system can predict if the patient is likely to have a certain disease or present incompatibility with some treatments. Considering the output of the classification model, the physician can make a better decision on the treatment to be applied to this patient [6].
The rest of the paper is organized as follows. The concept of Associative Classifiers (AC) is being discussed in section 2. In section 3 the advanced AC that are recently proposed have been introduced. In section 4 the refinement of support and confidence framework and its mathematical constraints ie downward closure property is discussed. In section 5 the future direction is given.

## 2. ASSOCIATIVE CLASSIFICATIONS
Associative classification(AC) mining is a promising approach in data mining that utilizes the association rule discovery techniques to construct classification systems. The entire dataset is divided into two part the 70% of data is used as training data and the remaining 30% is used for testing the accuracy of classifier. It is a three-step process shown in figure 1.

i. Generate the set of association rules from the training set with certain support and confidence thresholds as candidate rules.

ii. Pruning the set of discovered rules to weed out those rules that may introduce over fitting

iii. Classification Phase is the step to make a prediction for test data and measure the accuracy of the classifier.

Association rule mining is performed on transactional a database that contains a record of items purchased in a different transaction. An example of such data base is shown in Table 1. An

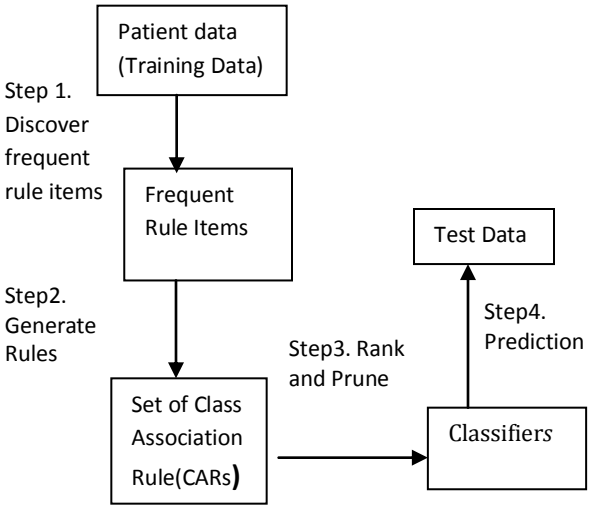


**Figure1: Associative Classifier for Data Mining**

**Table 1:Transactional Database**

| T | Items |
|---|---|
| 1 | A B C D E |
| 2 | A C E |
| 3 | B D |
| 4 | A D E |
| 5 | A B C D |

*association rule* is an implication of the form $A \Rightarrow B$, *where* $A, B \subseteq I$, *where I is set of all items, and* $A \cap B = \varnothing$. The rule $A \Rightarrow B$ has a *support s* in the transaction set *D* if *s*% of the transactions in *D* contain $A \cup B$ .The rule $A \Rightarrow$ B holds in the transaction set D with confidence c if c% of transactions in D that contain A also contain B . The problem of discovering all association rules from a set of transactions *D* consists of generating the rules that have a *support* and *confidence* greater than given thresholds. These rules are called *strong rules*, and the framework is known as the *support confidence framework* for association rule mining.

The AC is performed on relational databases that contains a records of entities of particular domain. Each record consist of m attribute $a_1$, $a_2$ …..$a_m$ and one important attribute known as predictive Attribute (Class Label). The attributes are continuous or categorical. Continuous attributes are required to be converted into categorical by using discritization technique. An example of such database from medical field is shown in table 2.
The AC is different from association rule mining Following are the terminology for AC analogous to Association Rule Mining.

i. Attribute which is pair (attribute, value), is used in place of Item. For example (BMI, 40) is an attribute in Table 2
ii. Attribute set is equivalent to Itemset for example ((Age, old), (BP, high)) .
iii. Support count of Attribute ($A_i$, $v_i$) is number of rows that matches Attribute in database.
iv. Support count of Attribute set ($A_i$, $v_i$)…($A_m$, $v_m$) is number of rows that matches Attribute set in data base.
v. An Attribute ($A_i$, $v_i$) passes the *minsup* threshold if

**Table 2. Sample Database for heart patient.**

| R_ID | Age | Smoking_habits | Hypertension | BMI | Heart_Disease |
|---|---|---|---|---|---|
| 1 | 42 | Yes | Yes | 40 | yes |
| 2 | 62 | Yes | No | 28 | No |
| 3 | 55 | No | Yes | 40 | yes |
| 4 | 62 | Yes | Yes | 50 | yes |
| 5 | 45 | No | Yes | 30 | No |

vi. An Attribute set (($A_i$, $v_i$)…($A_m$, $v_m$)) passes the *minsup* threshold if support count (($A_i$, $v_i$)…($A_j$, $v_j$)), ≥*minsup.*
vii. CAR Rules are of form (($A_i$, $v_i$), …,($A_j$, $v_j$))→c where c $\epsilon$ Class-Label. Where Left hand side is itemset and right hand sigh is class. And set of all attribute and class label together ie (($A_i$, $v_i$), …,($A_j$, $v_j$),c) is called rule attribute.
viii. Support count of rule attribute (($A_i$, $v_i$), …,($A_j$, $v_j$), c) is number of rows that matches item in database.
ix. Rule attribute (($A_i$, $v_i$), …,($A_j$ $v_j$), c) passes the *minsup* threshold if support count of (($A_i$, $v_i$), …,($A_j$, $v_j$), c) ≥*minsup* .

A special subset of association rules whose right-hand-side is restricted to the class attribute is used for classification. This subset of rules is referred as the Class Association Rules (CARs).Using association rule for classification is advantageous over traditional classifiers is that the simple if -then -else rules are used for classification which makes it easy for the end user to understand and interpret it. Also unlike decision tree approach one can easily update the rule set without affecting the complete rule set where as in decision tree approach it requires reshaping of complete tree

## 3. ADVANCEMENTS IN CAR RULE GENERATION.

The accuracy of AC largely depends on the set of rules we are having before classification will occur. If we are having Good Class Association Rules (CARs), definitely the accuracy will be high. Recently a number advanced Association Rule Mining techniques have been proposed to get good rules. These Advanced ARM can be combined with classifiers to give the Advanced AC having good predicting capabilities.

### 3.1 An associative Classifier based on positive and negative rules.

A new associative classifiers that take the advantage of negative association rule mining and associative classifiers. These are two relatively new domains of research. The paper extends the concept of positive association rule of the form $X \rightarrow Y$ to $\neg X \rightarrow Y$, $X \rightarrow \neg Y$ and $\neg X \rightarrow \neg Y$ with the meaning X is for presence and $\neg X$ is for absence. Instead of using support-confidence framework in the association rule generation the algorithm uses support-confidence some measure based on correlation analysis.

Correlation coefficient measure is added to support-confidence framework to find the interestingness of a rule. The correlation coefficient measures the strength of the linear relationship between a pair of two variables. For two variables X and Y it is given by $\rho = \frac{Con(X,Y)}{\sigma_X, \sigma_Y}$, where $Con(X,Y)$ represents the covariance of two variables and $\sigma_X$ stand for standard deviation. The range of values for σ is between -1 to +1. if the two variables are independent then σ equals to 0. When σ =+1 then variables considered are perfectly correlated. Similarly when σ = -1 then variables considered are perfectly negative correlated. The algorithm has been tested for UCI datasets and encouraging results are obtained when both positive and negative rules are used for classification. Also accuracy decreases when using only negative rules for classification. Negative association rules have been found effective to extract hidden knowledge useful information from medical databases.

## 3.2 Temporal Associative Classifiers

**T**he data is not always static in nature but changes with time and therefore adopting temporal dimension to this approach will give more realistic approach and will yield much better and useful results. The purpose of temporal predictive system is to provide the pattern or relationship among the items in time domain. For example rather than the basic association rule of {bread}→{butter} mining from the temporal data we can get a more insight rule that the support of {bread}→{butter} raises to 50% during 7 pm to 10 pm everyday [3]. These rules are more informative and useful to make a strategic decision making in every field. Temporal database is having an additional time related attribute.
Time is an important aspect of all real world phenomena.

- There are many examples of time-ordered data that demand our attention (*e.g.*, scientific, medical, dynamic systems, computer network traffic, web logs, markets, sales transactions, machine/device performance, weather/climate, telephone calls)
- The monitoring and tracking of real-world events frequently require repeated measurements – the volume of dynamic time-tagged data is therefore growing, and continuing to grow.
- Many of the data mining methods that we have studied require some modification to handle special temporal relationships ("before", "after", "during", "in summer", "whenever X happens")
- Time-ordered data lend themselves to prediction – what is the likelihood of an event, given the preceding history of events? (*e.g.*, hurricane tracking, disease epidemics)
- Time-ordered data often link certain events to specific patterns of temporal behavior (*e.g.*, network intrusion break-ins).

To deal with above such situation the new type of AC called Temporal Associative Classifier is being proposed in [3].The authors have Modified the three most popular AC ie CBA, CMAR and CPAR with temporal dimension and proposed TCBA, TCMAR and TCPAR. The authors have performed the experiment to compare the classifying accuracy and execution time of the three algorithm using temporally modified dataset of UCI machine learning data sets. The conclusion were

i. TCPAR performs better than TCMAR and is little better than TCBA as TCBA is time consuming for smaller support values and improves in run- time performance as the support increases.

ii. Using 10 dataset the accuracy is calculated for each algorithm. The average accuracy of TCPAR is found little better than TCMAR.

iii. The temporal counterpart of all the three associative classifiers has shown improved classification accuracy as compare to the non temporal associative classifier.

Time-ordered data lend themselves to prediction like what is the likelihood of an event, given the preceding history of events? *e.g.*, hurricane tracking, disease epidemics. The temporal data is useful in predicting the disease in different age group.

## 3.3 Associative Classifiers using Fuzzy Association Rule

In Classification problem the quantitative attributes are descritized as a one of preprocessing step. When the data are associated with quantitative domains such as income, age, price, etc., which are very common in many real applications, association rule mining usually needs to partition the domains in order to apply the Apriori-type method. Thus, a discovered rule $X \rightarrow Y$ reflects association between interval values of data items. Examples of suchrules are "*Fruit*[1-5kg] → *Meat*[5-20$]", "*Income*[20-50k$] → *Age*[20-30]", and so on[ ZC08] As a result the record belongs to only one of the set as result a sharp boundary problem. This gives rise to the notion of fuzzy association rules (FAR). The semantics of a fuzzy association rule is richer and of certain natural language nature, which are deemed desirable. For example, "*low-quantity Fruit* → *normal-consumption Meat*" and "*medium Income* → *young Age*" are fuzzy association rules, where X's and Y's are fuzzy sets with linguistic terms (i.e., *low*, *normal*, *medium*, and *young*). In the paper [ZC08] the authors have dealt with the "sharp boundary" problem for quantitative domains and have proposed an associative classification based on fuzzy association rules (namely CFAR).Fuzzy rules are found to be useful for prediction modeling system in medical domain. In medical domain most of the attributes are quantitative in nature hence fuzzy logic is used to deal with sharp boundary problems.

### *3.2.1 Defining Support and confidence measure:*

New formulae of support and confidence for fuzzy classification rule $F \rightarrow C$ is as follows:

$$\text{support}(F \rightarrow C) = \frac{\text{Sum of membership values of antecedent with class label C}}{\text{Total No. of Records in the Database}}$$

$$\textit{confidence}(F \rightarrow C) = \frac{\text{Sum of membership values of antecedent with class label C}}{\text{Sum of membership values of antecedent for all class label}}$$

For Example: Consider part of Database Part of a Database Containing Membership.

Table 3

| *M-Age* | *Class* |
|---|---|
| 0.1 | *C1* |
| 0.7 | *C1* |
| 0.8 | *C1* |
| 0.9 | *C2* |
| 0.5 | *C1* |
| 0.1 | *C2* |

suppose *F*= {*M-Age*} and *C*=*C1* and part of a database shown in Table 3. We have:

$$support(F \rightarrow C) = \frac{0.1 + 0.7 + 0.8 + 0.5}{6} = 35.0\%,$$

$$confidence(F \rightarrow C) = \frac{0.1 + 0.7 + 0.8 + 0.5}{0.1 + 0.7 + 0.8 + 0.9 + 0.5 + 0.1}$$

$$= 67.7\%.$$

## 3.4 Weighted Associative Classifiers

Weighted Associative Classifiers is another concept that assigns different weights to different features 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 predicting capability. A weighted associative classifiers consists of training dataset T={$r_1$, $r_2$, $r_3$…. $r_i$…} with set of weight associated with each {attribute, attribute value} pair. Each $i^{th}$ record $r_i$ is a set of attribute value and a weight $w_i$ attached to each attribute of $r_i$ tuple / record. In a weighted framework each record is set of triple {$a_i$, vi, $w_i$} where attribute $a_i$ is having value vi and weight wi, $0<w_j<=1$. Weight is used to show the importance of the item. Using Weighted Associative Classifiers, weighted rules like "*medium Income →young Age*" , "{(Age,">62"), (BMI,"45"), (Boold_pressur,"95-135")}, → Heart_Diseas , (Income[20,000-30,000]→Age[20-30]) can be used for classification. For the synthectic database given in table2, Weight of each record is calculated in Table 4, using Weight of different attribute in predicting the probability of Heart Disease from Table 5.

**Table 4. Relational Database with record weight.**

| R_ID | Age | Smoking_habits | Hypertension | BMI | Record weight |
|---|---|---|---|---|---|
| 1 | 42 | Yes | Yes | 40 | (0.2+0.8+0.6+0.8) /4=0.60 |
| 2 | 62 | Yes | No | 28 | (0.3+0.8+0.5+0.3) /4=0.42 |
| 3 | 55 | No | Yes | 40 | (0.2+0.8+0.6+0.5) /4=0.52 |
| 4 | 62 | Yes | Yes | 28 | (0.3+0.8+0.6+0.8) /4=0.67 |
| 5 | 45 | No | Yes | 30 | (0.2+0.7+0.6+0.3) /4=0.45 |

| S.No. | Symptoms | Weights |
|---|---|---|
| 1 | Age<40 | 0.1 |
| 2 | 40<age<58 | 0.2 |
| 3 | age>58 | 0.3 |
| 4 | Smoking_habits=yes | 0.8 |
| 5 | Smoking_habits=no | 0.7 |
| 6 | Hypertension=yes | 0.6 |
| 7 | Hypertension=no | 0.5 |
| 8 | BMI<=25 | 0.1 |
| 9 | 26<=BMI<=30 | 0.3 |
| 10 | 31<=BMI<=40 | 0.5 |
| 11 | BMI>=40 | 0.8 |

**Table 5 Weight of symptoms for heart disease (attribute weight).**

### *3.3.2 Defining Support and confidence measure:*

New formulae of support and confidence for classification rule X→Class_label, where X is set of weighted items, is as follows:

**Weighted Support:** Weighted support WSP of rule X→Class_label, where X is set of non empty subsets of attribute-value set, is fraction of weight of the record that contain above attribute-value set relative to the weight of all transactions.

$$WSP(X \rightarrow Class\_label) = \frac{\sum_{i=1}^{|X|} weight(r_i)}{\sum_{k=1}^{|n|} weight(r_i)}$$

**Example:** Consider a rule R (Hypertension="yes") → Heart_Disease="yes" then Weighted Support of R is calculated as:

$$WSP(R) = \frac{\text{Sum of Record Weight having the condition Hypertension='yes" true and also given class label Heart Disease}}{\text{Sum of Weight of all transactions}}$$

$$WSP(R) = \frac{0.42+0.40+0.50}{0.42+0.38+0.40+0.50+0.34}$$

$$WSP(R) = 0.640$$

**Definition 5. Weighted Confidence:** Weighted Confidence 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).

$$\text{Weighted Confidence} \quad \frac{\text{Weighted Support } (X \cup Y)}{\text{Weighted Support } (X)}$$

**Example:** The Weighted confidence of the rule R (Hypertension="yes") → Heart_Disease="yes" can be calculated as :

$$WC(R) = \frac{\text{Sum of Record Weight having the condition Hypertension='yes' true and also the class label Heart\_Disease}}{\text{Sum of Record Weight having the condition Hypertension="yes" true}}$$

$$WC(R) = \frac{0.42+0.40+0.50}{0.42+0.40+0.50+0.34}$$

$$WC(R) = 0.795$$

## 4. Refining Support and Confidence Measures to validate Downward closure property

The Apriori algorithm is based on the assumption that if the itemset is frequent, then all its subsets are also frequent. This allows the algorithm to build frequent itemsets of increasing size by adding items to itemsets that are already found to be frequent ie if the itemset is not frequent then its superset can't be frequent, this property is called downward closure property. The refinement of Support and Confidence used in any advanced Association rule mining should retain this property and an AC based on that advanced association rule miner too. In this WAC, instead of using support, a weighted support and weighted confidence , is used. The authors have proved that using weighted support the "weighted downward closure property" retains.

## 5. CONCLUSION & FUTURE WORK

In future these advanced AC can be combined to fulfil the real life requirements. For example Fuzzy and Weighted AC can be combined to give more accurate results as most of the data in medical is quantitative and weight can be applied to each feature based on their prediction capability. The temporal aspect can be used to improve the prediction for the patients of different age group.

## 6. REFERENCES

[1]. Sunita Soni ' Jyothi Pillai,O.P. Vyas An Associative Classifier Using Weighted Association Rule2009 International Symposium on Innovations in natural Computing, World Congress on Nature & Biologically Inspired Computing (NaBIC 2009), 978-1-4244-5612-3/09/$26.00 c_2009 IEEE 1492-1496

[2]. Zuoliang Chen, Guoqing Chen BUILDING AN ASSOCIATIVE CLASSIFIER BASED ON FUZZY ASSOCIATION RULES International Journal of Computational Intelligence Systems, Vol.1, No. 3 (August, 2008), 262 - 273

[3]. Ranjana Vyas, Lokesh Kumar Sharma, Om Prakash vyas, Simon Scheider Associtive Classifiers for Predictive analytics: Comparative Performance Study, second UKSIM European Symposium on Computer Modeling and Simulation 2008.

[4]. E.Ramaraj N.VenkatesanPositive and Negative Association Rule Analysis in Health Care Database, IJCSNS International Journal of Computer Science and Network Security, VOL.8 No.10, October 2008,325-330.

[5]. Khan, M.S. Muyeba, M. Coenen, F A *Weighted Utility Framework for Mining Association Rules*, Symposium Computer Modeling and Simulation, 2008. EMS '08. Second UKSIM European , page(s): 87-92.

[6]. Fadi Thabtah, *A review of associative classification mining*, The Knowledge Engineering Review**,** Volume 22, Issue 1 (March 2007), Pages 37-65, 2007.

[7]. Luiza Antonie, University of Alberta, Advancing Associative Classifiers - Challenges and Solutions, Workshop on Machine Learning, Theory, Applications, Experiences 2007

[8]. Feng Tao, Fionn Murtagh and Mohsen Farid. *Weighted Association Rule Mining using Weighted Support and Significance* Framework Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining 2003**,** Pages:661-666 Year of Publication: 2003

[9]. Yin, X. & Han, J. CPAR: Classification based on predictive association rule. In Proceedings of the SIAM International Conference on Data Mining. San Francisco, CA: SIAM Press, 2003,pp. 369-376.

[10]. W. Li, J. Han, and J. Pei. CMAR: Accurate and efficient classification based on multiple class association rules. In ICDM'01, pp. 369-376, San Jose, CA, Nov.2001.

[11]. Liu,B Hsu. W. Ma, Integrating Clasification and association rule mining . Proceeding of the KDD, 1998(CBA) pp 80-86.

[12]. Cláudia M. Antunes , and Arlindo L. Oliveira *Temporal Data Mining: an overview*.

[13]. Antonie, M. & Zaïane, O. **An associative classifier based on positive and negative rules.** In Proceedings of the 9th ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery, 2004, pp 64-69