Global Journal of Computer Science and Technology
Volume 11 Issue 22 Version 1.0 December 2011
Type: Double Blind Peer Reviewed International Research Journal
Publisher: Global Journals Inc. (USA)
Online ISSN: 0975-4172 & Print ISSN: 0975-4350

Data mining with Predictive analysis for healthcare sector: An Improved weighted associative classification approach

By Y.Shirisha, S.Siva Shankar Rao, D. Sujatha
Department of CSE

Abstract - Association mining has seen its growth right through data mining during the last few years as it has the ability to search for that entire database that could be of least constraints associated with it.Thus finding such small database sets could be done with the help of predictive analysis method. The paper enlightens the combinational classification of association and classification data mining. For this to happen a new set of constraints need to be introduced namely classification association rule(CAR).some systems like classification systems with domain experts are the ones that can be associated with. For fields like medicine where a lot many patients consult each doctor, but every patient has got different personal details not necessarily may suffer with same disease. So the doctor may look for a classifier, which could provide all details about every patient and henceforth necessary medications can be provided. However there have been many other classification methods like CMAR, CPAR MCAR and MMA and CBA.Some advance associative classifiers have also seen growth very recently with small amendments in terms of support and confidence, thereby accuracy. In this paper we proposed a HIT algorithm based automated weight calculation approach for weighted associative classifier.

Keywords : *classifier, Association rules, data mining, healthcare, Associative Classifiers, CBA, CMAR, CPAR, MCAR.*

GJCST Classification : *H.2.8*

Strictly as per the compliance and regulations of:

Data mining with Predictive analysis for healthcare sector: An Improved weighted associative classification approach

Y.Shirisha [α], S.Siva Shankar Rao [Ω], D. Sujatha [β]

Abstract - Association mining has seen its growth right through data mining during the last few years as it has the ability to search for that entire database that could be of least constraints associated with it.Thus finding such small database sets could be done with the help of predictive analysis method. The paper enlightens the combinational classification of association and classification data mining. For this to happen a new set of constraints need to be introduced namely classification association rule(CAR).some systems like classification systems with domain experts are the ones that can be associated with. For fields like medicine where a lot many patients consult each doctor, but every patient has got different personal details not necessarily may suffer with same disease. So the doctor may look for a classifier, which could provide all details about every patient and henceforth necessary medications can be provided. However there have been many other classification methods like CMAR, CPAR MCAR and MMA and CBA.Some advance associative classifiers have also seen growth very recently with small amendments in terms of support and confidence, thereby accuracy. In this paper we proposed a HIT algorithm based automated weight calculation approach for weighted associative classifier.

Keywords : *classifier, Association rules, data mining, healthcare, Associative Classifiers, CBA, CMAR, CPAR, MCAR.*

I. Introduction

A set of steps followed to extract data from the related pattern is termed as data mining. From a haphazard data set it is possible to obtain new data. The predictive modeling approach that simply combines association and classification mining together [3] shows better accuracy [10]. The classification techniques CBA [10], CMAR [9], CPAR [8] out beat the traditional classifiers C4.5, FOIL, RIPPER which are faster but not accurate. Associate classifiers are fit to those model applications which provide support domains in the decisions. However the most suited example for this is medical field where in the data for each patient is required to be stored, with the help of which the system predicts the diseases likely to be affecting the patient. With the system throughput the doctor may decide the medication [6].

Author α : M.Tech, Department of CSE, ATRI, Parvathapur, Uppal, Hyderabad, India. E-mail : shirisha37@gmail.com
Author Ω : Assoc.Prof. Department of CSE, ATRI, Parvathapur, Uppal, Hyderabad, India. E-mail : sivasankarssr@gmail.com
Author β: Assoc.Prof and HOD Department of CSE, ATRI, Parvathapur, Uppal, Hyderabad, India. E-mail: sujatha.dandu@gmail.com

II. Association Classification

This method of mining is an algorithm based technique with support of some association constraints to retrace the data. The data available is fragmented into 2 parts.one of which is 70% of total data and is taken as training data and the rest taken as another part is used for testing purpose. This technique of data mining is done on data base sets with a record of information.

- Step1: with the help of training set of data produce an association rule set.
- Step2: eliminate all those rules that may cause over fitting.
- Step3: finally we predict the data and check for accuracy and this is said to be the classification phase.

One such example of data base set is:

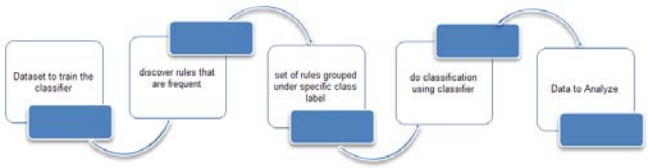


Figure 1 : Associative Classifier for Data Mining

Transaction ID	Items in Transaction
1	ABCDE
2	ACE
3	BD
4	ADE
5	ABCDE

Table 1 : Transactional Database

An association rule is an implication of the form $A \Rightarrow B$, where $A, B \subseteq I$, where I is set of all items, and $A \cap B = \phi$. The rule $A \Rightarrow B$ has a support s in the transaction set D if s% of the transactions in D contain $A \bigcup B$. 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. if the threshold point is crossed in terms of confidence then the association rules could be determined. Thus the determined rules form a confidence frame with the help of high strength rules.

On a particular set of domains the AC is performed. A tuple is a collection of m attributes $a_1, a_2, ..., a_m$ and consists of another special predictive Attribute (Class Label). However these attributes can be categorized depending action methods. Association rule mining is quite different from AC but still undergo following similarities:

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 match Attribute set in data base.
v. An Attribute (A_i, v_i) passes the minsup threshold if support count $(A_i, v_i) \geq \min \sup$.

Record ID	Age	Smokes	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

Table 2 : Sample Database for heart patient.

vi. An Attribute set $((A_i, v_i)....(A_m, v_m))$ passes the threshold if support count $((A_i, v_i), (A_{i+1}, v_{i+1})....(A_j, v_j)) \geq \min \sup$.
vii. CAR Rules are of form where c∈ $((A_i, v_i), (A_{i+1}, v_{i+1})....(A_j, v_j)) \rightarrow c$ Class-Label. Where Left hand side is itemset and right hand sigh is class. And set of all attribute and class label together ie ((Ai, vi), …,(Aj, vj),c) is called rule attribute.
viii. Support count of rule attribute $((A_i, v_i), (A_{i+1}, v_{i+1})....(A_j, v_j), c)$ is number of rows that matches item in database. Rule attribute $((A_i, v_i), (A_{i+1}, v_{i+1})....(A_j, v_j), c)$ passes the threshold if support count of $((A_i, v_i), (A_{i+1}, v_{i+1})....(A_j, v_j), c) \geq \min \sup$.

An important subset of rules called class association rules (CARs) are used for classification purpose since their right hand side is used for attributes. Its simplicity and accuracy makes it efficient and friendly for end user. Whenever any amendments need to be done in a tree they can be made without affecting the other attributes.

III. Advancements in Car Rule Generation

The accuracy of the classification however depends on the rules implied in the classification.To overcome CARs rules inaccuracy in some cases, a new advanced ARM in association with classifiers has been developed. This new advanced technique provides high accuracy and also improves prediction capabilities.

a) An Associative Classifier Based On Positive And Negative Approach

Negative association rule mining and associative classifiers are two relatively new domains of research, as the new associative amplifiers that take advantage of it. The 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. Based on correlation analysis the algorithm uses support confidence Instead of using support-confidence framework in the association rule generation. Correlation coefficient measure is added to support confidence framework as it measures the strength of 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 σx stand for standard deviation.

The range of values for σ is between -1 to +1, when it is +1 the variables are perfectly correlated, if it is -1 the variables are perfectly independent then equals to 0. when positive and negative rules are used for classification in UCI data sets encouraging results will obtain. Negative association rules are effective to extract hidden knowledge. And if they are only used for classification, accuracy decreases.

b) Temporal associative classifiers

As data is not always static in nature, it changes with time, so adopting temporal dimension to this will give more realistic approach and yields much better results as the purpose is to provide the pattern or relationship among the items in time domain. For example rather than the basic association rule of $\{\text{bread}\} \rightarrow \{\text{butter}\}$ mining from the temporal data we can get that the support of $\{\text{bread}\} \rightarrow \{\text{butter}\}$ raises to 50% during 7 pm to 10 pm everyday [3],as These rules are more informative they are used to make a strategic decision making. Time is an important aspect in temporal database.

1. Scientific, medical, dynamic systems, computer network traffic, web logs, markets, sales, transactions, machine/device performance, weather/climate, telephone calls are examples of time ordered data.
2. The volume of dynamic time-tagged data is therefore growing, and continuing to grow,
3. as the monitoring and tracking of real-world events frequently require repeated measurements.3.Data mining methods require some modification to handle special temporal relationships ("before", "after", "during", "in summer", "whenever X happens").
4. 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)
5. Time-ordered data often link certain events to specific patterns of temporal behavior (e.g., network intrusion breaks INS).The new type of AC called Temporal Associative Classifier is being proposed in [3] to deal with above such situation. CBA, CMAR and CPAR are modified with temporal dimension and proposed TCBA, TCMAR and TCPAR. To compare the classifying accuracy and execution time of the three algorithms using temporally modified data set of UCI machine learning data sets an experiment has performed and conclusions are:

   i. TCPAR performs better than TCMAR and TCBA as it is time consuming for smaller support values but improves in run- time performance as the support increases.
   ii. Using data set 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 e.g., (hurricane tracking, disease epidemics). The temporal data is useful in predicting the disease in different age group.

c) Associative Classifier Using Fuzzy

Association Rule: The quantitative attributes are one of preprocessing step in classification. for the data which is associated with quantitative domains such as income, age, price, etc., in order to apply the Apriori-type method association rule mining needs to partition the domains. Thus, a discovered rule $X \rightarrow Y$ reflects association between interval values of data items. Examples of such rules are "Fruit [1-5kg] $\rightarrow$ Meat [5-20$]", "Income [20-50k$] $\rightarrow$ Age [20-30]", and so on [ZC08]. As the record belongs to only one of the set results in sharp boundary problem which gives rise to the notion of fuzzy association rules (FAR).The semantics of a fuzzy association rule is richer and natural language nature, which are deemed desirable. For example, "low-quantity Fruit $\rightarrow$ normal-consumption Meat" and "medium Income $\rightarrow$ 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).An associative classification based on fuzzy association rules (namely CFAR) is proposed to overcome the "sharp boundary" problem for quantitative domains. Fuzzy rules are found to be useful for prediction modeling system in medical domain as most of the attributes are quantitative in nature hence fuzzy logic is used to deal with sharp boundary problems.

i. *Defining Support and confidence measure*

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

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

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

d) Weighted Associative Classifiers

Based on the different features weights are allotted based on this classifier. Every attribute varies in terms of importance.it also important to know that with the capabilities of predicting the weights may be altered. A weighted associative classifiers consists of training dataset T={r1,r2, r3…. ri…} with set of weight associated with each {attribute, attribute value} pair. Each with recordri is a set of attribute value and a weight wi attached to each attribute of rituple / record. Aweighted framework has record as a triple {ai, vi, wi} where attribute ai is having value vi and weight wi, 0<wj<=1. Thus with the help of weights one can easily determine its predicting ability. With this weighted rules like "medium Income young Age", "{(Age,">62"), (BMI,"45"), (Boold_pressur,"95-135")}, Heart Disease, (Income[20,000-30,000]Age[20-30]) could become the criteria of determination. Weights of data as per table 2 are recorded in table 4 with the help of weights mentioned in the table 5 for predicting attributes.

Record ID	Age	Smokes	Hypertension	BMI	Record Weight
1	42	YES	YES	40	0.6
2	62	YES	NO	28	0.42
3	55	NO	YES	40	0.52
4	62	YES	YES	50	0.67
5	45	NO	YES	30	0.45

Table 4 : Relational Database with record weight

Here we measure the record weight using fallowing

$$recordWeight = \frac{\sum_{i=1}^{|I|} w_i}{|I|}$$

Here

$|I|$ is total number of items in record

i is an item in record

w_i is weight of the item i

i. *Measuring weights using HITS algorithm*

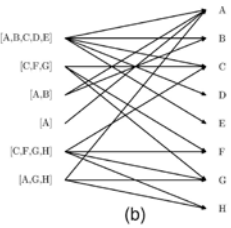
a. Ranking Transactions with HITS

A database of transactions can be depicted as a bipartite graph without loss of information. Let $D = \{T_1, T_2, ...T_m\}$ be a list of transactions and $I = \{i_1, i_2, ...i_n\}$ be the corresponding set of items. Then, clearly D is equivalent to the bipartite graph $G = (D, I, E)$ where

$$E = \{(T, i) : i \in T, T \in D, i \in I\}$$

TID	Transaction
100	[A, B, C, D, E]
200	[C, F, G]
300	[A, B]
400	[A]
500	[C, F, G, H]
600	[A, G, H]

(a)



(b)

Fig1 : The bipartite graph representation of a database (a) Database (b) Bipartite graph

Example 1: Consider the database shown in Fig. 1a. It can be equivalently represented as a bipartite graph, as shown in Fig. 1b. The graph representation of the transaction database is inspiring. It gives us the idea of applying link-based ranking models to the evaluation of transactions. In this bipartite graph, the support of an item i is proportional to its degree, which shows again that the classical support does not consider the difference between transactions. However, it is crucial to have different weights for different transactions in order to reflect their different importance. The evaluation of item sets should be derived from these weights. Here comes the question of how to acquire weights in a database with only binary attributes. Intuitively, a good transaction, which is highly weighted, should contain many good items; at the same time, a good item should be contained by many good transactions. The reinforcing relationship of transactions and items is just like the relationship between hubs and authorities in the HITS model [3]. Regarding the transactions as "pure" hubs and the items as "pure" authorities, we can apply HITS to this bipartite graph. The following equations are used in iterations:

$$auth(i) = \sum_{T:i \in T} hub(T),\ hub(T) = \sum_{i:i \in T} auth(i) ....(1)$$

When the HITS model eventually converges, the hub weights of all transactions are obtained. These weights represent the potential of transactions to contain high-value items. A transaction with few items may still be a good hub if all component items are top ranked. Conversely, a transaction with many ordinary items may have a low hub weight.

b. W-support - A New Measurement

Item set evaluation by support in classical association rule mining [1] is based on counting. In this section, we will introduce a link-based measure called w-support and formulate association rule mining in terms of this new concept.

The previous section has demonstrated the application of the HITS algorithm [3] to the ranking of the transactions. As the iteration converges, the authority weight $auth(i) = \sum_{T:i \in T} hub(T)$ represents the "significance" of an item I, accordingly, we generalize the formula of $auth(i)$ to depict the significance of an arbitrary item set, as the following definition shows:

Definition 1: The w-support of an item set X is defined as

$$w\sup p(X) = \frac{\sum_{T:X \subset T \wedge T \in D} hub(T)}{\sum_{T:T \in D} hub(T)} ....(2)$$

Where $hub(T)$ is the hub weight of transaction T. An item set is said to be significant if its w-support is larger than a user specified value Observe that replacing all $hub(T)$ with 1 on the right hand side of (2) gives `supp(X)`. Therefore, w-support can be regarded as a generalization of support, which takes the weights of transactions into account. These weights are not determined by assigning values to items but the global link structure of the database. This is why we call w-support link based. Moreover, we claim that w-support is more reasonable than counting-based measurement.

ID	Symptoms	Weight
1	Age<40	0.437
2	40<Age<58	0.375
3	Age>58	0.185
4	Smokes=yes	0.68
5	Smokes=no	0.31
6	Hypertension=yes	0.5
7	Hypertension=no	0.5
8	BMI<=25	0.23
9	26<=BMI<=30	0.2
10	31<=BMI<=40	0.43
11	BMI>40	0.125

Table 5 : weights measured using proposed algorithm

ii. *Defining Support and confidence measure*

New formulae of support and confidence for classification rule $X \rightarrow class\_Label$, where X is set of weighted items, is as follows: Weighted Support: Weighted support WSP of rule $X \rightarrow class\_Label$, 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.

$$v1 = \sum_{i=1}^{|X|} weight(r_i)$$

$$v2 = \sum_{k=1}^{|n|} weight(r_i)$$

$$WSP(X \rightarrow Class\_Label) = \frac{v1}{v2}$$

The classification accuracy improvement for Weighted associative classifier with proposed weight measurement approach can be observable in fig 2 and fig 3.

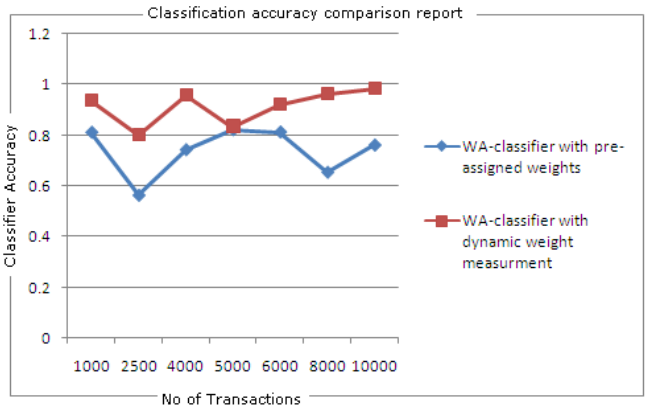


Fig 2 : A line chart representation of classification accuracy differences between Traditional Weighted associative classifier and proposed weighted associative classifier

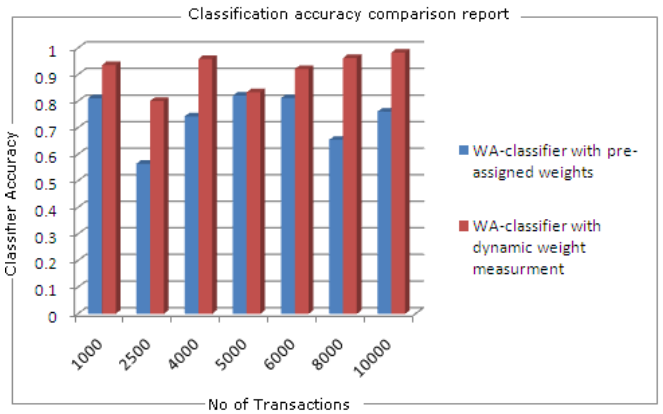


Fig 3 : A bar chart representation of classification accuracy differences between Traditional Weighted associative classifier and proposed weighted associative classifier

IV. Refining Support and Confidence Measures to Validate Downward Closure Property

The downward closure property is the key part of Apriori algorithm.it states that any super set can't be frequent unless and until its itemset isn't frequent. The itemsets that are already found to be frequent are added with new items based on the algorithm. However changes in support and confidence shall not show its effect on this property and also AC associated with advanced rule developer. The terms support and confidence are to be replaced with weighted support and weighted confidence respectively in WAC which elicits that weighted support helps maintain weighted closure property.

V. Conclusion

This advanced AC method could be applied in real time scenario to get more accurate results. This needs lot of prediction to be done based on its capabilities which could be improved.it find it major application in the field of medical where every data has an associated weight. The proposed HIT algorithm based weight measurement model is significantly improving the quality of classifier.

References References Referencias

1. SunitaSoni, JyothiPillai, O.P. Vyas An Associative Classifier Using Weighted Association Rule2009 International Symposium on Innovations in natural Computing, World Congress on Nature & Biologically.
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. RanjanaVyas, Lokesh Kumar Sharma, Om Prakashvyas, Simon ScheiderAssocitive Classifiers for Predictive analytics: Comparative Performance Study, second UKSIM European Symposium on Computer Modeling and Simulation 2008.
4. E.RamarajN.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. FadiThabtah, A review of associative classification mining, The Knowledge Engineering Review, Volume 22, Issue 1 (March 2007), Pages 37-65, 2007.
7. LuizaAntonie, University of Alberta, Advancing Associative Classifiers - Challenges and Solutions, Workshop on Machine Learning, Theory, Applications, Experiences 2007
8. Feng Tao, FionnMurtagh 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.