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

Y.SHIRISHA[1], S.SIVA SHANKAR RAO[2] and D. SUJATHA[3]

[1] AURORAS TECHNOLOGICAL AND RESEARCH INSTITUTE/JNTU

Received: 27 October 2011 Accepted: 22 November 2011 Published: 2 December 2011

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.

Index terms— classifier, Association rules, data mining, healthcare, Associative Classifiers, CBA, CMAR, CPAR, MCAR.

1 INTRODUCTION

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].

? 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: 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 iii. Support count of Attribute (,)i i

A v is number of rows that matches Attribute in database.

iv. Support count of Attribute set (,),...., (,)i i m m A v A v

is number of rows that match Attribute set in data base.

v. An Attribute (,)i i A v passes the minsup threshold if support count (,) min sup i i A v ? .

Table **??** : Sample Database for heart patient.

vi. An Attribute set ((,)....(,))i i m m A v A v passes the threshold if support count**1 1**

((,), (,)....(,)) min supi i i i j j A v A v A v + + ? .

vii. CAR Rules are of form where c?1 1

((,), (,)....(,))i i i i j j A v A v A v c + + ? Class-Label.

Where i i i i j j A v A v A v c + +

is number of rows that matches item in database. Rule1 1

((,), (,)....(,),)i i i i j j A v A v A v c + + passes the threshold if support count of**1 1**

((,), (,)....(,),) min supi i i i j j A v A v A v c + + ? .

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.

2 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.

3 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 formX Y ? to X Y ¬ ? , X Y ? ¬ and X Y ¬ ? ¬ with the meaning X is for presence and 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 (,) ,con X Y x y ? ? ? =

, where (,) con X Y represents the covariance of two variables and x ? stand for standard deviation.

The range of values for? 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.

4 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 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 nontemporal 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.

5 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 Aprioritype method association rule mining needs to partition the domains. Thus, a discovered rule X ? Y reflects association between interval values of data items.

Examples of such rules are ”Fruit [1-5kg] ? Meat [5-20$]”, ”Income [20-50k$] ? 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 ? 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).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. 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 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:: : ( ) ( ), ( ) ( )....(**1**) T i T i i T auth i hub T hub T auth i ? ? = = ? ?

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.

6 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 -. [1] [2] [3] [4] [5]

1

Figure 1: Figure 1 :

[1] © 2011 Global Journals Inc. (US) Global Journal of Computer Science and Technology Volume XI Issue XXII Version I

[2] December

[3] © 2011 Global Journals Inc. (US) Global Journal of Computer Science and Technology Volume XI Issue XXII Version I 33 2011 December

[4] © 2011 Global Journals Inc. (US) Global Journal of Computer Science and Technology Volume XI Issue XXII Version I 34

[5] © 2011 Global Journals Inc. (US) Global Journal of Computer Science and Technology Volume XI Issue XXII Version I 35 2011 December

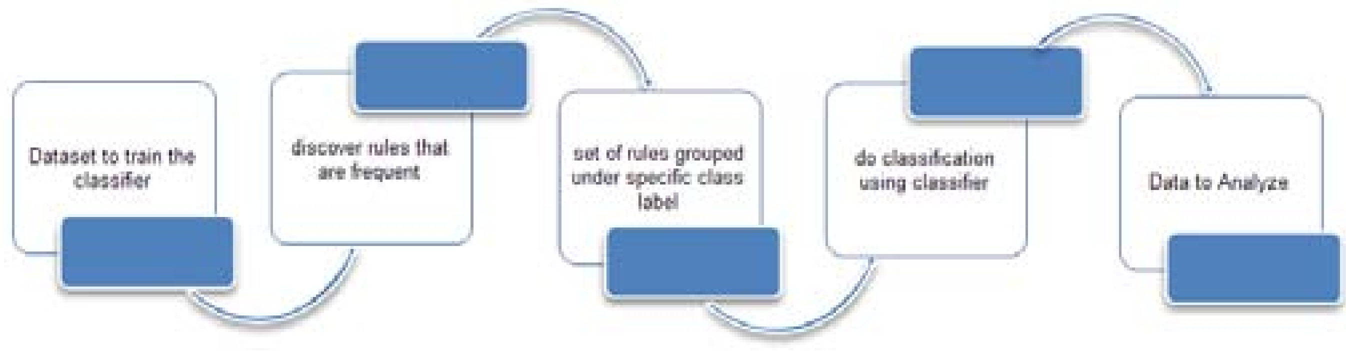
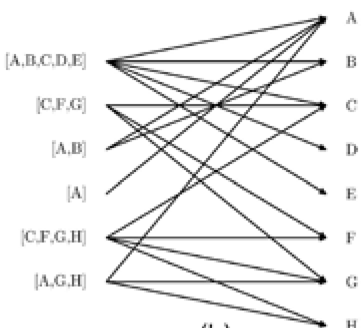


Figure 2:

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)

Figure 3: Fig1:

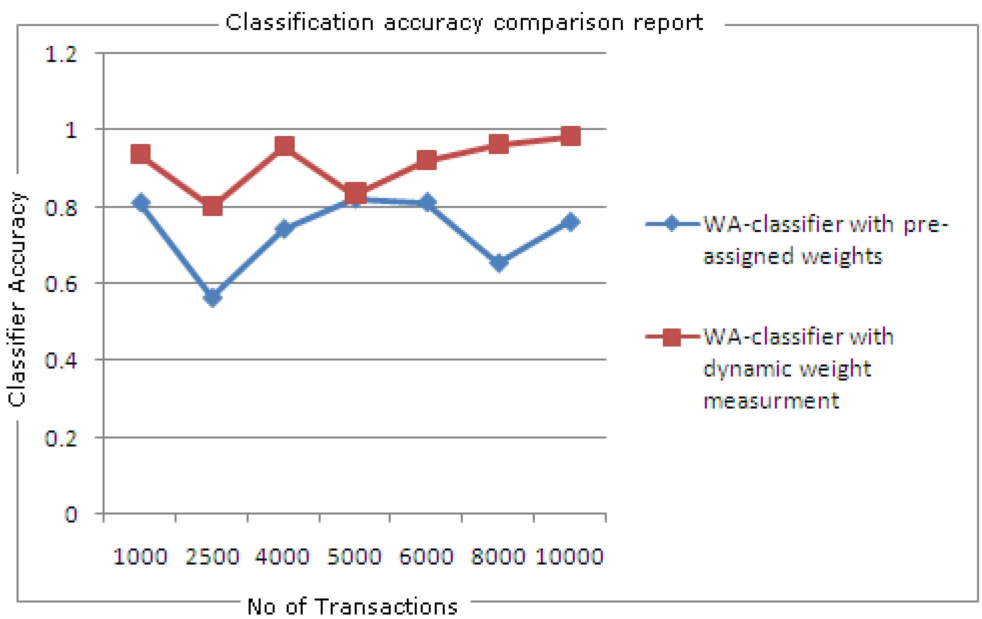
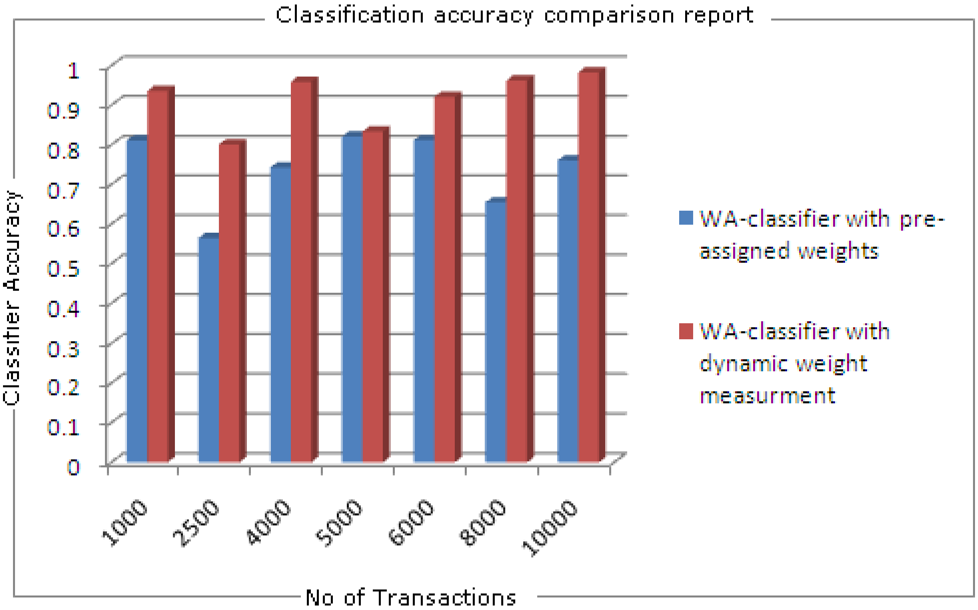


Figure 4:



3

Figure 5: Fig 3 :

sup Sum of membership values of antecedent with class label C ? =) port F C (

Total No. of Records in the Database

Sum of membership values of antecedent with class label C ? =) confidence F C (

Sum of membership values of antecedent for all class label

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

Figure 6:

4

F ? C

Figure 7: Table 4 :

T. An item set is said to be significant if its w-support is larger than a user specified value Observe that replacing all 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 wsupport link based. Moreover, we claim that w-support is more reasonable than counting-based measurement.

.1 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.

.2 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.

[Chen (2008)] , Zuoliang Chen . *Guoqing Chen BUILDING AN ASSOCIATIVE CLASSIFIER BASED ON FUZZY ASSOCIATION RULES International Journal of Computational Intelligence Systems* August, 2008. 1 (3) p. 273.

[IJCSNS International Journal of Computer Science and Network Security (2008)] , *IJCSNS International Journal of Computer Science and Network Security* October 2008. 8 (10) p. .

[Fadithabtah (2007)] 'A review of associative classification mining'. Fadithabtah . *The Knowledge Engineering Review* March 2007. 2007. 22 (1) p. .

[Luizaantonie and Of ()] 'Advancing Associative Classifiers -Challenges and Solutions'. University Luizaantonie , Alberta Of . *Workshop on Machine Learning, Theory, Applications*, 2007.

[Antonie and zaïane ()] 'An associative classifier based on positive and negative rules'. M Antonie , O &zaïane . *Proceedings of the 9th ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery*, (the 9th ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery) 2004. p. .

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

[Li et al. (2001)] 'CMAR: Accurate and efficient classification based on multiple class association rules'. W Li , J Han , J Pei . *ICDM'01*, (San Jose, CA) Nov.2001. p. .

[Yin and Han ()] 'CPAR: Classification based on predictive association rule'. X Yin , J Han . *Proceedings of the SIAM International Conference on Data Mining*, (the SIAM International Conference on Data MiningSan Francisco, CA) 2003. SIAM Press. p. .

[Liu et al.] 'Integrating Clasification and association rule mining'. Liu , W Hsu , Ma . *Proceeding of the KDD, 1998(CBA)*, (eeding of the KDD, 1998(CBA)) p. .

[Sunitasoni] Jyothipillai Sunitasoni , OP . *Vyas An Associative Classifier Using Weighted Association Rule2009 International Symposium on Innovations in natural Computing, World Congress on Nature & Biologically*,

[Khan et al. ()] *Utility Framework for Mining Association Rules, Symposium Computer Modeling and Simulation*, M S Khan , M Muyeba , Coenen . 2008. p. . (EMS '08. Second UKSIM European, page(s)

[Tao and Farid ()] 'Weighted Association Rule Mining using Weighted Support and Significance Framework'. Feng Tao , Fionnmurtagh , Mohsen Farid . *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, (the ninth ACM SIGKDD international conference on Knowledge discovery and data mining) 2003. 2003. p. .