Emerging Trends in Associative Classification Data Mining

Neda Abdelhamid and Aladdin Ayesh
Computing and Informatics, Leicester, UK
Email: p09050665@myemail.dmu.ac.uk, ayesh@dmu.ac.uk

Fadi Thabtah
Ebusiness Dept., Canadian University of Dubai, Dubai, UAE
Email: fadi@cud.ac.ae

Abstract—**Utilising association rule discovery to learn classifiers in data mining is known as Associative Classification (AC). In the last decade, AC algorithms proved to be effective in devising high accurate classification systems from various types of supervised data sets. Yet, there are new emerging trends and that can further enhance the performance of current AC methods or necessitate the development of new methods. This paper sheds the light on four possible new research trends within AC that could enhance the predictive performance of the classifier or their quality in terms of rules. These possible research directions are considered starting research points for other scholars in rule based classification in data mining.**

Index Terms—**artificial intelligence, associative classification, classification, research trends**

I. Introduction

Two data mining tasks specifically classification and association rule are correlated in which association rule finds relationships among attribute values in a database whereas classification's goal is allocating class labels to test data [1], [2]. When these tasks are merged the result is Associative Classification (AC) which employs association rule to only discover the rules and adds on top of that additional steps, i.e. (sorting, pruning, and prediction) [3].

Normally, an AC algorithm operates in three main phases. During the first phase, it looks for hidden correlations among the attribute values and the class in the input data and generates them as "Class Association Rule" (CARs) in "IF-THEN" format [4]. After the complete set of CARs are found, ranking and pruning procedures (phase 2) start operating where the ranking procedure sorts rules according to certain thresholds such as confidence and support. Further, during pruning, useless rules are discarded from the complete set of CARs. The output of phase 2 is the set of CARs which represents the classifier. Lastly, the classifier derived gets tested on new independent data set to measure its effectiveness in forecasting the class of unseen test cases.

Manuscript received October 11, 2013; revised April 16, 2014.

Research studies for instance [5] have shown that AC mining has two distinguishing features over other traditional classification approaches. The first one is that it produces very simple knowledge (rules) that can be easily interpreted and manually updated by the end-user. Secondly, this approach often finds additional useful hidden information and therefore the error rate of the resulting classifier is minimised. The motivation of this article is to highlight future research directions and emerging trends related to AC. These can be considered as starting points that can be addressed by future research works in data mining to further enhance the performance of current AC algorithms.

The rest of the chapter is organised as follows: The AC problem and its solution scheme are discussed in Section 2. Section 3 is devoted to the different new trends and possible research directions in AC. Finally conclusions are demonstrated in Section 4.

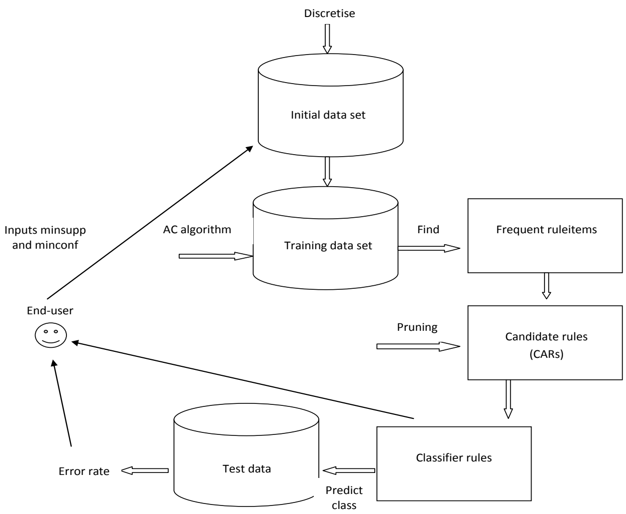

Figure 1. The general life cycle of the AC algorithm

II. Associative Classification

AC in data mining is rule based learning approach that joins association rule and classification together. This approach had come to surface as a promising research discipline in a paper titled "Integrating classification and

doi: 10.12720/ijeee.3.1.50-53

association rule" [6]. In AC, the training phase (rule learning) is about searching for hidden knowledge among the attribute values and the class and then the classifier is constructed after sorting the knowledge and pruning redundant knowledge. Many research studies, i.e. [7], [8], revealed that AC usually extracts good classifiers with reference to error rate. The AC algorithm life cycle is depicted in Fig. 1.

A. The AC Problem Statement

Given a training data set *D*, which has *n* distinct attributes A_1, A_2, …, A_n and *C* is a list of classes. The number of cases in *D* is denoted |*D*|. An attribute may be categorical (where each attribute takes a value from a known set of possible values) or continuous. For categorical attributes, all possible values are mapped to a set of positive integers. In the case of continuous attributes, any discretisation method can be applied. The goal is to construct a classifier from *D*, e.g. $Cl: A \rightarrow C$, which can forecast the class of test cases where *A* is the set of attribute values and C is the set of classes.

The majority of AC algorithms mainly depend on a threshold called *minsupp* which represents the frequency of the attribute value and its associated class (AttributeValue, class) in the training data set from the size of that data set. Any attribute value plus its related class that passes *minsupp* is known as a frequent ruleitem, and when the frequent ruleitem belongs to a single attribute, it is said to be a frequent 1-ruleitem. Another important threshold in AC is the *minconf*, which can be defined as the frequency of the attribute value and its related class in the training data set from the frequency of the attributes value in the training data.

III. New Trends and Future AC Research Problems

A. Immune Systems Based AC

One of the effective learning approaches that has been originated from the Natural Immune System (NIS) and have successfully applied in optimization, online security and data mining is Artificial Immune System (AIS). As a matter fact, AIS has been utilised in classification problem in last decade and devised a competitive performance results in accuracy rate. Examples of known classification algorithms that are based on AIS are clonal selection and negative selection. We believe that AIS can be used in AC especially to minimise the search space for rules by reducing the number of candidate rules. Hereunder, two attempts in using AIS within AC have been outlined.

There have been some initial attempts to adapt the learning methodology of NIS especially the clonal selection in AC context that have resulted in an algorithm named artificial immune system-associative classification (AIS-AC). The AIS-AC algorithm was proposed in 2005 and extended in 2009 and follows the evolutionary process by reducing the search space of the candidate rules by keeping just high predictive rules. This process is accomplished by extracting frequent 1-ruleitems after passing over the initial training data set, and generating the possible candidate ruleitems at iteration *N* from results derived at iteration *N* - 1 and so forth. The *minsupp* and *minconf* are utilised as sharp lines to discriminate among rulesitems at each iteration. Further, two new parameters are introduced named Clonal_rate and Max_generation. The clonal_rate (defined below) denotes the rate at which items in the candidate rules at given generation are extended, and it is proportional to the rule confidence.

$$Clonal\_rate = \frac{n.Clonal\_rate}{\sum_{i=1}^{n} conf(r_i)} \quad (1)$$

where *n* is the number of rules at the current iteration, and the clonal_rate is a predefined user parameter. Once the candidate rules are extracted, they are tested on the training data keeping only those that have one or more training example(s) coverage. The algorithm terminates once the complete training data set is covered or the Max_generation condition has been met (often set to 10). The candidate rules that have training data coverage are kept in the classifier. The AIS-AC algorithm applies the rules in the classifier on the test data similar to CBA prediction method.

Recently, another AIS based on AC called AC-CS was proposed in [5]. This algorithm follows the same track of the previously described AIS-AC and it uses the same strategies in deriving the rules and classifying test data. One simple difference between AC-CS and AIS-AC is that AC-CS builds the candidate rules in generations per class rather than at once and then merges each class rules set before evaluating the complete set of rules on the training data to determine the classifier.

Empirical evaluations using a limited number of UCI data sets indicated that the AIS algorithm is highly competitive in accuracy and execution time to the "Predictive Apriori" algorithm which is a simplified version of CBA that primarily uses Apriori algorithm for extracting the rules without pruning.

B. Test Data Training

Lazy AC as an approach was originated to maximize the predictive power of classifiers by minimising rule filtering to only candidate rules that wrongly cover training data while building the classifier. Recently, [9], [10] have proposed a new lazy approach in AC mining that primarily depends on the test data attribute values in reducing the rules set applicable to the test data in the classification step. Hereunder, we briefly shed the light on two different lazy learning methodologies and introduce an important issue in classification related to delaying learning rules until the classification step. This can be seen as a possible research starting point to minimise the search space for candidate rules.

The first learning methodology in lazy AC focuses on minimizing candidate rules filtering process aiming to accomplish high performance classifiers in regards to accuracy rate. Precisely, lazy AC algorithms that follow this methodology like L^3G discard only candidate rules that have wrong classification when evaluating rules in the process of building the classifier. Meaning, while evaluating rules on the training data to choose the best ones, all rules that either a) have correct classification on the training data or b) have not covered any training data, are stored in the classifier. The rules that correctly cover at least one training data are stored in a primary storage, and the rules that have no training data coverage are kept in a lower secondary storage. These two storages together are simply the classifier. Now, when the classification process of test data starts, rules in the primary storage are checked and when none of them is able to classify the test data rules in the secondary storage are utilised instead of the default class rule. This approach normally produces very large classifiers which may restrict its utilisation for applications.

A different approach was proposed in [9], [10] which allows both training and testing examples to play a role in assigning the class to the test data. This learning methodology claims that deriving all candidate rules in the training phase could be problematic in cases when the *minsupp* is set to low values. Thus, suggesting using the test data attribute values as valuable information to reduce the search space of applicable rules. This reduces the dimensionality of the training data though it requires learning from part of the training data and for each test data repeatedly similar to Naïve bayes even if data are not partitioned with respect to class labels as in Naïve bayes algorithm.

C. Calibration

Accuracy is one of the main metrics used in classification algorithms in data mining to favour an algorithm over others for certain data sets. In fact, most of classification problems such as credit card scoring, website classification, weather forecasting, etc., use accuracy or its complement one-error-rate as the main evaluation metric to distinguish among classification algorithms. Though, certain applications like cost-sensitive classification, Information Retrieval ranking in search engines, and text categorization for digital libraries, may require additional information beside classification accuracy such as class membership probabilities per test. So in calibrated AC approach, the derived rules per test data are used to describe the training data set and these rules are utilised to compute the class membership probabilities. When the rules are accurate calibrated AC algorithms assumes that the estimated class membership probabilities are also accurate and can be generalised.

There are many classic rule based and non-rule based approaches in classification that have employed calibration. Some of which are SVM, decision trees, and probabilistic. In AC, one calibrated approach has been used AC, i.e. [10]. We believe that calibration is an important issue that should be studied extensively in AC simply since initial results revealed good predictive performance if compared to other current algorithms. Furthermore, for multiple label classification including the class membership probabilities are much more useful than single label classification because of two reasons. Firstly, in multi-label classification, the input data instance may belong to several classes and therefore we can assign weights or class memberships in particular when classes overlap in the training data. Thus, the decision maker can distinguish easily to which the input data belongs to or can merge multiple classes together to come up with new class label. Secondly, some of the rules in the classifier will be connected to set of classes and therefore calibration can assist in prioritising these classes (Ranking).

D. Non Confidence Based Learning

The key element, which controls the number of rules produced in AC is the support threshold. If the support is set to a large value, normally the number of extracted rules is very limited, and many rules with high confidence will be missed. This may lead to discarding important knowledge that could be useful in the classification step. To overcome this problem, one has to set the support threshold to a very small value. However, this usually involves the generation of massive number of classification rules, where many of which are useless since they hold low support and confidence values. This large number of rules may cause severe problems such as overfitting.

[11], [12] argued that the rule confidence which is the main criteria for selecting the classifier could be misleading in some cases especially since the rule with the largest confidence is chosen to predict the test case in the test data set. So, instead of computing the confidence from the training data set as most AC methods, the test data should be considered in favouring rules during the prediction phase. Therefore, the authors proposed a measure of rule goodness called “predictive confidence” which is based on statistical information in the test data set (the frequencies of the test cases applicable to a rule). The new predictive confidence based AC approach is called AC-S. This approach is required to calculate the rule (R) “confidence decrease” = R(Conf(Training)) – R(Conf(Testing)) in order to estimate the predictive confidence for each rule before predicting test cases.

The AC-S algorithm depends on several parameters that must be known at the time of prediction and for each test case before the algorithm chooses the most applicable rule to the test case. Precisely, the support and confidence for each candidate rule must be computed and from both the training and testing data sets so that AC-S can be able to estimate the predictive accuracy for each rule. This indeed is time consuming and can be a burden in circumstances where the training data set is highly correlated. Further, it is impractical to estimate the support and confidence for each rule in the testing data set in advance since we don’t know which rule will be used for prediction. Yet, we can utilise the test data during the prediction step to narrow down candidate rules. This can be seen a new research path for enhancing the current “predictive confidence” approach. A comparison between AC-S and other known AC algorithms such as CBA, CBA (2) and CMAR was conducted against some UCI data sets. The results of the accuracy showed that AC-S is competitive to CBA, though CBA (2) and CMAR algorithms derived higher quality classifiers than AC-S.

IV. CONCLUSIONS

Several AC algorithms have been developed in the research literature in the last few years where most of which outperform other rule based classification approaches such as decision trees and rule induction. Nevertheless and since the fast development in the area of rule based classification new research problems rise which require taking care of by researchers in the same domain. In this paper, we highlighted four important research trends and applications that may need to be treated by AC algorithms to further enhance the performance of current algorithms. Primarily, this article addressed Artificial Immune AC, Calibration AC, Learning from the test data set, and non-support based learning. We believe that these areas require careful consideration and be researched further by scholars in rule based classification.

REFERENCES

[1] N. Abdelhamid, A. Ayesh, F. Thabtah, S. Ahmadi, and W. Hadi, "MAC: A multiclass associative classification algorithm," *Journal of Information and Knowledge Management (JIKM)*, vol. 11, no. 2, pp. 1250011-1 - 1250011-10, 2012.

[2] N. Abdelhamid, A. Ayesh, and F. Thabtah, "An experimental study of three different rule ranking formulas in associative classification mining," in *Proc. 7th IEEE International Conference for Internet Technology and Secured Transactions (ICITST)*, 2012, pp. 795-800.

[3] E. Baralis, S. Chiusano, and P. Graza, "A lazy approach to associative classification," *IEEE Transactions on Knowledge and Data Engineering*, vol. 20, no. 2, 2008.

[4] G. Costa, R. Ortale, and E. Ritacco, "X-Class: Associative classification of XML documents by structure," *ACM Trans. Information Syst*ems, vol. 31, no. 1, pp. 3, 2013.

[5] S. Elsayed, S. Rajasekaran, and R. Ammar, "AC-CS: An immune-inspired associative classification algorithm," *ICARIS*, 2012, pp. 139-151.

[6] B. Liu, W. Hsu, and Y. Ma, "Integrating classification and association rule mining," in *Proc. Knowledge Discovery and Data* Mining *Conference-KDD*, New York, 1998, pp. 80-86.

[7] F. Thabtah, Q. Mahmood, L. mccluskey, and H. Abdel-Jaber, "A new classification based on association algorithm," *Journal of Information and Knowledge Management*, vol. 9, no. 1, pp. 55-64, 2010.

[8] F. Thabtah, P. Cowling, and Y. Peng, "A study of predictive accuracy for four associative classifiers," *Journal of Digital Information Management*, vol. 3, no. 3, 2005.

[9] A. Veloso, W. Meira, M. Zaki, M. Goncalves, and H. Mossri, "Calibrated lazy associative classification," *Information Sciences: An International Journal*, vol. 13, no. 181, pp. 2656-2670, 2011.

[10] A. Veloso, W. Meira, M. Gonçalves, and M. Zaki, "Multi-Label lazy associative classification," in *Proc. of Principles of Data Mining and Knowledge Discovery*, 2007, pp. 605-612.

[11] C. H. Wu, J. Y. Wang, and C. J. Chen, "Mining condensed rules for associative classification," *ICMLC*, vol. 4, pp. 1565-1570, 2012.

[12] X. Xu, G. Han, and H. Min, "A novel algorithm for associative classification of images blocks," in *Proc. the fourth IEEE International Conference on Computer and Information Technology*, 2004, pp. 46-51.

Neda Abdelhamid is a final year PhD student in the Computer Science and Informatics department at De Montfort University, UK. Her research project involves developing data mining models for single and multi-label classification in data mining. She had seven publications in reputable journals and international conferences and she is expected to do her PhD defense in spring, 2014.

Aladdin Ayesh – MSc (Essex, 1995), PhD (LJMU, 2000), has a Chartered status in engineering (CEng) and as an IT practitioner (CITP). He is a Fellow of British Computer Society (FBCS), a Fellow of Higher Academy of Education (FHAE), and a Senior Member of IEEE (SMIEEE). Dr. Ayesh is Reader in Artificial Intelligence at De Montfort University. His research involves theoretical and practical work in the areas of swarm intelligence, cognitive systems and natural language processing. Dr. Ayesh is an editor for 4 journals and numerous conferences, and has over 100 published works in international journals, conferences, edited and sole-authored books, and professional magazines.

Fadi Thabtah is an Associate Professor in e-business at the Canadian University of Dubai. His research interests lie in the investigation and development of new data mining, scheduling and machine learning techniques, which bridge the gap between the theory and practice of decision making, using artificial intelligence techniques. Dr. Fadi participated in several technical committees board for international journals and conferences, i.e. Chair for the IEEE ITNG '08 conference, AI-05, AI-06, AI-07, DBA-06, Journal of Information System, International Journal of Software Engineering and Knowledge Engineering, Journal of Applied Soft Computing, Journal of Software Engineering and Knowledge Engineering, etc. He is currently supervising 4 PhD students working in Data Mining, and previously had successfully supervised 9 PhD and 12 Mphil students. He had published over 70 journal and conference articles in data mining, big data, phishing, and computer networks.