International Journal of Engineering & Technology, 7 (4) (2018) 5659-5665

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET
doi: 10.14419/ijet.v7i4.28379
Research paper

RCuA: rule classification use association data mining model for structure and unstructured data

Mohammed Hayel Refai [1] *, Saleh Ali Alomari [2], Tarik Khalil [1], Hussam Saleh Abu Karaki [3]

[1] Department of Information Systems and Technology, Sur University College, Sur, Oman
[2] Faculty of Science and Information Technology, Jadara University, Irbid, Jordan
[3] Faculty of Information Technology at Al Hussein Bin Talal University, Ma'an, Jordan
**Corresponding author E-mail: moahmmedhayel@suc.edu.om*

Abstract

Association and classification rule mining are important activities in the data mining domain. Incorporating the association rule discovery and classification within this domain leads to a method, called the associative classification method. Text Categorizations (TC) prevails form major problems through this domain including machine learning communities. This issue is not simple to be solved since available data has enormous dimensionality. There exist particular enormous amounts of online documents within a group of data in which each data is combined along with a particular class. Categorization refers to a structure of design from a categorized data, which categorizes past unrecognized documents as accurate as it could be. The paper proposes a novel text classification model by applying an Associative Classification (AC) model, namely, the Rule Classification use Association (RC*u*A), which produces an obvious text document. Additionally, the paper attempts at forming an expansion of available AC of current associative text classifiers, which cope with structure and unstructured English document assemblies. The produced model is tested through two experiments of structure and unstructured data. The first experiment is related to the UCI datasets, while the second is related to Reuters-21578 datasets. The experiment is based on utilizing various classification categorization learning algorithms (e.g. MCAR and CBA) in order to assess the efficiency of the proposed model in this paper. As a result, it is found to be proven from the findings that the new RC*u*A model improves the accuracy of the dataset in comparison with the MCAR and CBA algorithms where the number of existing rules is decreased. The RC*u*A makes an average accuracy of 83.945% compared to the CBA and MCAR algorithms resulting with an accuracy of 82.34% and 83.655%, respectively. In terms of unstructured dataset, the RC*u*A produces an average accuracy of 89.328% in comparison with the CBA and MCAR algorithms resulting with an accuracy of 77.34% and 83.64286%, respectively.

Keywords: *Text Categorization; Associative Classification; CBA; MCAR; RCuA; UCI; Reuters-21578.*

1. Introduction

Data mining is considered to be a major problem within the domain of Knowledge Discovery Database (KDD) process. It includes the use of data analysis and discovery algorithms for producing specific facts of models (i.e. patterns) within an involved data. This production is based on adequate computational effectiveness restraints [1]. The remaining stages of the KDD include data cleansing, pattern validation, data reduction, data selection and visualization that are all related to the explored information [2]. In the data mining domain, one of the major applied tasks refers to the classification task. The major aim of classification is to produce a model from a set of characteristics such that each characteristic refers to a goal class [3], [4]. This model is utilized to predict the classes of a new group of attributes [5], [6]. Classification is implemented through various domains, such as in medical analysis domain, space exploration [7] and text classification [8].

Text Classification (TC) has been one of the popular task in text mining [9-11] where it includes the understanding, organization and recognition of many different kinds of textual data [12]. The main aim of the TC is to categorize an inward textual document within a group. The "supervised" learning classification classifies a current document over a prearranged input text assembly [4]. The TC forms a multi-stage procedure, which processes textual documents by classifying the document according to an algorithm and by assessing the created classification model [13]. Different categorization models are utilized in the TC where these are implemented and being used through the domains of data mining and machine learning, such as decision trees [14], Naive bayes [15], [16], Support Vector Machine [17], [18] and neural network [19]. These models are basically tested and utilized when categorizing English documents [11]. However, a model of the TC called the Association Classification (AC) [20] denotes the research domain in data mining by combining an association rule discovery along with the categorization process [21], [22]. The major aim of the AC is based on creating a model, which is called the classifier [23], [24]. This model contains a particular quantity of knowledge from a labeled input, with an intention of envisaging a class characteristic for a test data case, which as precise as possible [25].

In the past years, many AC algorithms were improved (e.g. Classification Passed Association Rules (CPAR) [32], Live and Let Live (L3G) [25], Multi Class Association Rule MCAR [26], CACA [27], BCAR [28], LCA [21] and many other algorithms). Earlier researchers indicate that the AC method creates precise classifiers in comparison with other data mining methods (e.g. probabilistic and decision tree). Unlike some conventional data mining approaches, such as probabilistic approaches and neural network approaches, which create classification methods that are difficult

Copyright © 2018 Mohammed Hayel Refai et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

to be comprehended or explained by an end-user, the AC method creates instructions that are easily comprehended and handled by all end-users [29]. There exist many new produced AC methods (e.g. CAEP [30], CMAR [31], CPAR [32], MCAR [26], CACA [27], ACCF and BCAR, CBA, Negative-Rules [30], Negative-Rules [33] and MMAC [34]. Such methods utilize various methodological and discovery rules, ranking rules, prune rules and prediction rules.

Associative Classification receives a high quantity of consideration from many different research societies. Nonetheless, it is seen that there exist very few experimentations, which are performed by utilizing the scope of a text categorization with just a small number of associative classification algorithms. Furthermore, these algorithms are able to overcome many text categorization issues that are being encountered [35]. A detailed review and a comprehensive analysis pertaining to these algorithms that are related to the text categorization discipline play an important role in this paper, including effective and distinguished associative classification algorithms that are identified through other main data mining domains.

In particular, there exist a few emphases that are based on improved performance methods and algorithm adjustments or mixtures. Such emphases are selected to be applied in later stages of this research. Such methods show that they are found the most effective methods within particular disciplines that are taken into account for adaptation in order to assist some parts of the algorithm basis baseline that are utilized for comparative findings. A comprehensive test for the current association rule mining algorithms are carried out through the current stage, representing important performance developments within the rule exploration procedure. This test is detailed and taken into account for possible combination via the associative classification algorithm. In particular, the combination of effective associative rule exploration algorithms is capable of providing higher computed velocities and reducing memory needs. However, the classification accurateness is not impacted since the output of the entire associative rule mining algorithms must be similar to each other. This paper is aimed at investigating the ability of applying the associative classification data mining model for an automatic classification of structure and unstructured texts of English documents.

2. Related research

The AC model is considered to be a method within the data mining domain that is efficiently being utilized to tackle real categorization problems (e.g. in image processing [36], medical diagnoses [31] and bioinformatics [21]. A number of empirical researches demonstrate that this technique is extremely precise in structuring a classification model [23], [24]. Additionally, the AC model produces rules that are not likely to be identified when utilizing conventional classification algorithms. This model creates "If-Then" rules that are more understandable and manageable for end-users [24]. Moreover, the ability to apply the AC classification method is basically the result of many advantages that are obtained by this method such as the ease of forming a classification model (classifier), a high prediction correctness and the easily preserved classifier in which rules are simply organized, inserted and detached [32].

The AC model can also be defined as a common classification method, which is based on an association rule. However, this model receives great attraction of a from various researchers. Its algorithms are utilized within a slight and an average sized dataset of the machine learning Irvine (UCI) at the University of California [37]. The application of the AC model that is based on a classifier of text assemblies is still under investigation. Hence, one of decisive aims regarding the investigation is based on using this model within organized and unstructured content build-ups. Although there exist a number of interests through various disciplines, very few investigations are focused to be used within the territory of text classification. Additionally, only a few number of different associative classification algorithms are available and are capable of addressing the problems of text classification. The AC approach may need a few numbers of user-specified attributes, which are based on the use of an associative rule mining approach. Common associative classification approaches indicate to a major problem in which a considerable number of rules are created [32]. This number is unreasonable as it revokes the procedure of utilizing the created rules, which are restricted in real-world implementations (e.g. text classification). Since the number of rules that is created by the algorithm reaches several thousands of rules [39] and could be replicated and indistinct through different required classes, they are required to get pruned based on different pruning processes. The rules that are left after pruning is being mere are required in order to shape a design that can be applied or a classified. Consequently, they are applied to particular categories of a current data. Nonetheless, when the number of rule is minimized, the classification accurateness is also minimized.

Bing at 1998 [40] were the first to involve the association rule algorithms including its classification. Within their research, two-stage processes are conducted such that the first stage identifies the repeated set of an item that is applying the Apriori Algorithm [41]. On the other hand, the classifier is built in the second stage. Through their testing, it is seen that their approach creates findings, which are compared to different common classification approaches (e.g. the decision tree).

The CMAR algorithm that is produced represents a further example of the AC approaches that is not based on one rule [31]. Nonetheless, it rather chooses and investigates the correlation within these rules that are of an increased confidence score. A tree named 'the CR-tree' is constructed to maintain particular rules that are kept in a descending order. This tree is followed by an attribute frequency through the antecedent section of this rule. Another AC approach comprises the Multi Class Association Rule (MCAR) as studied in Thabtah, et al. (2005). In fact, this approach involves a ranking rule method where just the rules with an increased level of confidence are maintained to be utilized within the classification. A new AC method, namely, the Multi Class Association Rule (MCAR) is introduced in [26].

The Modified Multi-class Classification based on Association Rule (MMCAR) is defined as a method for an association classification that uses a rule productive function for minimizing the number of rules by applying a pruning process, namely, the Partly Rule Match (PRM) [38]. An earlier research that investigates the use of the association rule through the classification benchmarks refers to the Classification Based on Association (CBA) rule. The CBA rule applies the Apriori algorithm [41] for exploring the repeated rule elements. This phase is named the candidate generation phase. The repeated rule elements comprise the <attributes, values> and a class that surpasses the minsupp. After that, the repeated rule elements are utilized to create a full group of CARs that are applied to design a classifier. This phase is named as the classifier building phase. Through the data collection phase, the data is gathered for two experimentations. In the first experimentation, the data group uses 14 UCI data elements [37]. In the second experimentation, the maximum occupied classifications of the Reuters-21578 [42] test gathering are applied.

3. The rule classification use association model (RCuA)

The Association Classification (AC) is considered a talented data mining method that constructs further precise classifiers in comparison to what the conventional classification method constructs. The Rule Classification use Association (RCuA) model contains five major stages as shown in Fig 2. The first stage of this method as seen from the design in Fig 1 involves the data pre-processing and representation. In this stage, a group of data involves the procedure of pre-processing that contains the tokenization, stop word removal and stemming procedures. The Tokenization indicates to the procedure of splitting a sequence of words within a sentence

into separated tokens. The Stop word removal refers to the procedure of deleting words, which are called, 'determiners' and 'prepositions' that are known in representing meaningless words of the learning algorithm. The stemming procedure refers to the procedure of modifying words through to their root as shown in Figure 1.

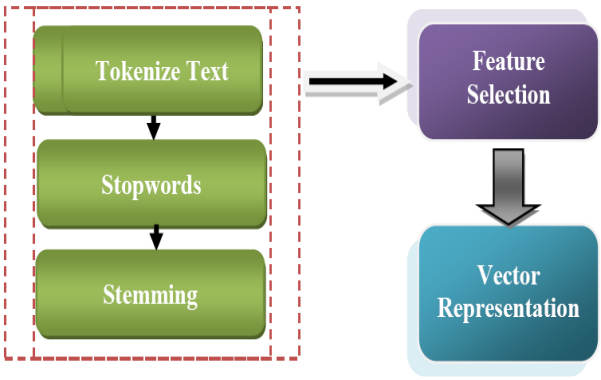


Fig. 1: The Pre-processing Operation in Text Mining.

The second stage of this method refers to the data representation. A mixture of horizontal and vertical design data layout is applied in the produced model of this paper. In this model, an element is embodied by a line number in which the first element is encountered through a group of data and through a column number for an element. This implies that every element is switched into (ColumnId, RowId) illustration that forms simple integers. Consequently, searching for elements for computing confidence and support values through the rule discovery procedure is rapid and needs less memory. According to this illustration, every first emergence of a word within a text is given a distinctive integer identification number. When the second stage is completed, the task is to fulfill unstructured data and make it structural. Table 1 illustrates an example of an input group of data. Table 2 shows its illustration.

Table 1: Examples of an Item That Is Explored within Every Line

TiD	Items
1	sea, port , wind
2	port, aqaba
3	port, corn
4	sea, port, aqaba
5	sea, corn
6	port, corn
7	sea, corn
8	sea, port, corn, wind
9	sea, port, corn

Table 2: An Item Illustration

TiD	Items ids
1	(1)1,(2)1,(3)1
2	(2)1,(4)2
3	(2)1,(5)3
4	(1)1,(2)1,(4)2
5	(1)1,(5)3
6	(2)1,(5)3
7	(1)1,(5)3
8	(1)1,(2)1,(5)3,(3)1
9	(1)1,(2)1,(5)3

The least initialization that is performed for the data is based on putting a distinctive integer value for every operation. The line numbers are used as a Transaction ID (TID). After that, elements are charted to its integer illustration (item ids) where each element is substituted with integer values of two sections, which comprise: Column Ids and row Id Ex (column1 "sea", row 1 it represents (1) 1). When the algorithm needs to compute support regarding an element when identifying whether it is repeated or not, an accumulation function is appealed to every group element and counts their presences through a data structure. In the produced model, a mixture of vector representation and term frequency are used in this paper in order to switch the increased dimensionality regarding a collection of Reuter text of a matrix. Additionally, the model computes the repeated elements by utilizing simple TID list connections. The Term Frequency (TF) is applied to calculate the importance of the keyword and its contributions to the output classifier [43]. The TF is considered to be one of the term weighting approaches that calculate the frequency of a keyword within a text where it is given in the Equation below:

$$tf\ (f_j, d_j) = \frac{freq_{ij}}{\max_k freq_{kj}}$$

When characteristics are selected by utilizing the TF, the following phase includes the shift of a text within an ordinary numerical procedure, which is appropriate for many different learning algorithms (structured data). This is known as a middle point such that a text data shifts through to a conventional data mining encoding process [44].

The third stage of this method refers to the identification of frequent set and rule discovery. Through this stage, the input dataset is scanned to explore the repeated elements within the form <Attribute-Value, class> of size 1. These elements are named as one-items. After that, the algorithm connects them frequently to crate frequent two-items, and so forth. It is important to note that any item appearing within the input dataset is lower than the MinSupp threshold and must be ignored.

Repeated elements utilize an intersection approach according to the Tid-list in order to calculate the confidence and support values regarding the characteristic values (elements). The Tid-list of an element shows the row numbers through the training dataset of where this element has happened. Therefore, among the intersection pertaining to the Tid-lists of the two disjoint elements, the resulting set refers to the row numbers where the current resulting item appears within the training dataset, and the cardinality of the resulted set illustrates the current element supporting value. This effective model computes the support value of the entire elements without the need to frequently scan the training dataset.

The first stage is to identify the elements that are thought to be repeated. In this paper, the repeated elements are illustrated via the intersection model according to the Tid-list [45],to compute the support and confidence values for the rule items having size greater than one are computed. For instance, for a class of item sets with prefix x, the following formula is expressed as:

[x] = {a1, a2, a3, a4}

The intersection of xai is performed with the entire xaj with j>i in order to obtain new classes. From [x], we can get classes as follows:

[xa1] = {a2, a3, a4}, [xa2] = {a3, a4}, [xa3] = {a4}

Therefore, when the entire repeated rule items recognize the confident rules as a second stage, the confidence value for every item set is greater than the lowest confidence (MinConf) threshold where one rule is only generated. This rule embodies itself as: X→ C, where C denotes the class of the greatest frequency for an Item set X.

These refer to the so-called frequent one-items, which are stored in a vertical manner. Nevertheless, the elements are taken apart if they are not greater than the lowest supporting value (MinSupp). After that, the candidate two-item is generated based on the use of the Tid-lists for the frequent one-item. This is conducted by crossing the Tid-lists of any two disjointed one-items. The procedure is performed frequently in order to produce the repeated three elements and the following repeated elements. When using this method, the only rules which are embodied in a statistical procedure and which contain high numbers of confidence values are created.

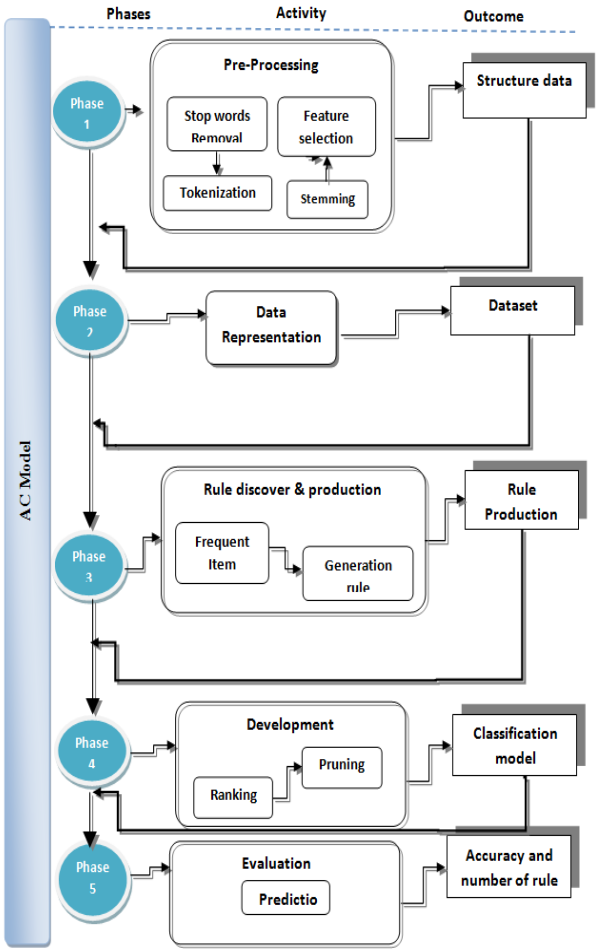


Fig. 2: The RuAC Model for Structured and Unstructured Data.

The fourth stage of this method refers to the development rule ranking of the AC that hails as a significant step, and which assists in selecting the most efficient rules for estimation. Each rule-based AC method conducts a sorting procedure through the rules within the procedure of a classifier-building. The sorting procedure forms the initial step of pruning noisy and useless rules. Prior to the performance of the prune repeated rules including the construction of the classifier, rules should be organized to more effective priority of the maximum quality rule representing a portion of the classifier. The rule ranking assists in pruning and overlying particular rules with less certainty in comparison with the broad rules. The rule pruning is considered an important step in AC mining approaches, that involves removing unimportant rules, and which causes an imprecise estimation [46], [47]. This step is normally applied once the entire rules are explored and organized such that a process or multiple processes are retrieved to prune repeated rules. For every reorganised rule (candidate classifier rule), the algorithm starts with an initial rule and its ability to be applied and assessed against the training situation. The rule is added through to the classifier once it partly copes with at least a single training situation regardless of the class conclusiveness.

The fifth stage of this model refers to the evaluation process. Based on the estimation of the testing data situation, the rule estimation splits the entire rules that are matching the test situation into different sets for every class, and it measures the support and confidence values pertaining to every set. Lastly, it allocates the testing situation and the class of the set, which include the biggest confidence. In such situations where there exist two or multiple sets with a nearly similar degree of confidence, the estimation approach is anticipated on the biggest support set.

The improvement of designing the study relies on the performed theoretical research. The whole performance of the research is illustrated within this section, which involves data pre-processing and illustration, the identification of repeated sets and rule discovery, rules determination, structure classification and assessment that copes with both structured and unstructured data.

4. Empirical analysis and findings

This paper introduces the findings that are gained through conducting different tests of structured and unstructured data. The experiment on structured data test is conducted for the UCI datasets and various classification learning algorithms are used (MCAR and CBA) to assess the efficiency of the RCuA model. The experiment on unstructured data test is conducted for the Reuters-21578 datasets by utilizing various classification learning algorithms (MCAR and CBA) in order to assess the efficiency of the RCuA model. The experiments are carried out for structured and unstructured data where the initial tests use 10 UCI datasets that are considered the most common applied benchmark for experimental assessment and present learning algorithms. The second tests use the 7 most occupied classifications of the Reuters-21578 datasets. In this paper, a cross assessment that splits the training dataset into (n+1) folds arbitrary is employed where rules are learned from n folds within every repetition and then evaluated on the remaining hold out fold. Furthermore, the major factors of the tests comprise the MinSupp and MinConf, which are set to 2% and 50% respectively, through the conducted tests. Table 3 represents the UCI datasets that are utilized within these tests.

Table 3: A Summary of the UCI Datasets

Data set	Size	No. of Classes
Austra	690	2
Balance-scale	625	3
Breast	699	2
Glass	214	7
Heart-s	294	2
Iris	150	3
Labor	57	2
Led7	3200	10
Lymph	148	4
Mushroom	8124	2

According to the studies conducted in the literature [48] regarding the text mining field, the most popularly used dataset refers to the Reuters-21578. Documents of the Reuters-21578 collection appear on the Reuters newswire and are indexed by personnel. This research needs a Reuters-21578 version ModApte including 9,174 documents in total. The data is separated by a particular expert into 2,579 documents for experimental purposes and 6,630 training documents. Table 4 depicts the number of training documents and testing groups for each classification of REUTERS-21578.

Table 4: The Number of Training and Testing Documents in REUTERS-21578

Category	Training	Testing
Acq	1650	719
Crude	389	189
Earn	2877	1078
Grain	433	149
Interest	347	130
Money-FX	538	197
Trade	396	117

Table 5 highlights the findings of the classification accurateness and the number of rules, which are created by the RCuA, MCAR and CBA and by utilising the structured dataset of the UCI.

Table 5: The Accurateness and the Number of Rules Regarding the UCI Dataset

	Accuracy			Number of rule		
Dataset	CBA	MCAR	RCuA	CBA	MCAR	RCuA
Austra	85.4	86.14	86.26	121	193	163
Balance-scale	68.2	76.96	76.17	45	77	19
Breast	94.7	94.99	93.83	61	79	60
Glass	69.9	71.35	74.2	36	43	27
Heart-s	71.2	81.15	80.51	35	39	36
Iris	93.3	92.93	94.26	16	16	11

Labor	95.0	83.5	83.5	17	15	15
Led7	72.4	71.83	73.2	53	214	83
Lymph	74.4	78.10	78.05	38	54	34
Mushroom	98.9	99.60	99.67	38	42	42

An analysis regarding the number of rules that are derived by the classifier is carried out. Figures 3 and 4 illustrate the size of the classifier that is explored for every UCI dataset based on the use of the RCuA, MCAR and CBA models.

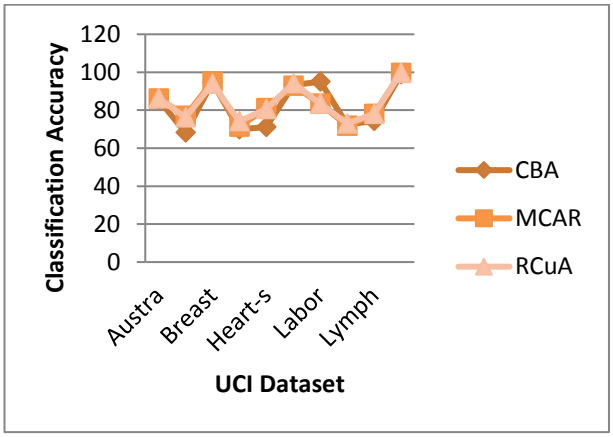


Fig. 3: The Classification Accurateness of the UCI Datasets.

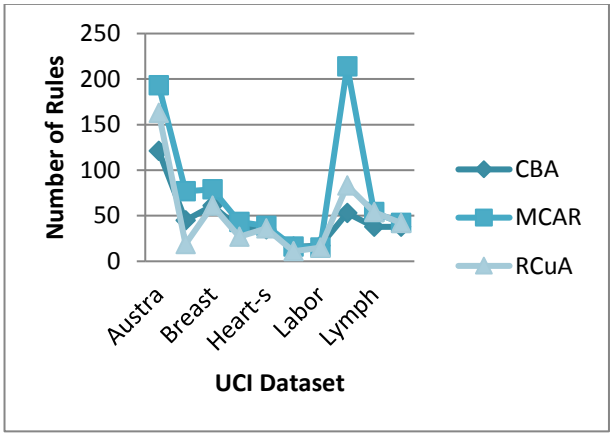


Fig. 4: The Number of Rules of the UCI Datasets.

A comparison between the CBA and RCuA has proven that the findings of the RCuA perform more effective results within 6 datasets compared to the CBA. A comparison between both of them has proven from the findings related to the RCuA that more effective results of 8 datasets are achieved in comparison with the findings in the MCAR. These findings refer to the fact that the RCuA algorithm obtains less number of rules through the majority of the situations in comparison with other particular algorithms.

A comparison between the two algorithms has proven that the findings from the RCuA algorithm are more precise for 8 datasets in comparison with the findings in the CBA algorithm that shows more precise results within the Labour and Breast pertaining to the datasets. A comparison between the MCAR and RCuA has proven that the findings from the RCuA are more precise for 5 datasets in comparison with the findings in the MCAR algorithm that shows more precise results within the Balance-Scale, Breast, Heart-s, Mushroom and Lymph pertaining to the datasets. Furthermore, the findings are seen to be the same in the labor dataset.

The RCuA algorithm performs an effective accuracy over the entire datasets. The RCuA algorithm achieves more precise findings in comparison with the CBA and MCAR algorithms regarding the majority of datasets. The findings illustrate that the RCuA algorithm improves the accuracy for 8 datasets in comparison with the CBA algorithm and minimizes the number of rules for 6 datasets. On the other hand, the RCuA algorithm improves the accuracy for 5 datasets more than the MCAR algorithm and minimizes the number of rules for 8 datasets. The connections between the classification accuracy and the number of rules for every UCI dataset refer to a positive linear connection. Table 6 highlights the findings of the classification accuracy and the number of rules that are created via the RCuA, MCAR and CBA algorithms based on the use of the unstructured dataset REUTERS-21578.

Table 6: The Accuracy and the Number of Rules for the REUTERS-21578 Dataset

	Accuracy			Number of rules		
Data	CBA	MCAR	RCuA	CBA	MCAR	RCuA
Acq	89.9	90.2	98.4	27	27	16
Crude	77	88.1	81.7	4	4	3
Earn	89.2	99.8	98.4	17	17	17
Grain	72.1	95.3	98.5	5	5	3
Interest	70.1	41.6	59.2	2	2	3
Money-FX	72.4	74.3	93.2	12	12	16
Trade	69.7	96.2	96.9	6	6	6

The RCuA performs more effective findings for 3 datasets once the CBA algorithm performs more effective findings for 2 datasets (Interest and Money-FX). Meantime, the findings are the same in Trade and Earn. Additionally, the RCuA algorithm achieves more effective findings for 3 datasets, while the MCAR performs more effective findings for a single dataset in Interest. At the same time, the findings are seen to be the same in Trade and Earn. The RCuA algorithm performs more effective findings for 6 datasets, while the CBA algorithm performs more effective findings for a single dataset in Interest. The RCuA algorithm performs more effective findings for 5 datasets, while the MCAR algorithm performs more effective findings for two datasets in Earn and Crude. In conclusion, the RCuA algorithm performs more effective accuracy for the entire datasets. It is found to be proven from the findings that the RCuA algorithm improves the accuracy for 6 datasets more than the CBA algorithm and minimises the number of rules for three datasets. On the contrary, the RCuA improves the accuracy for 5 datasets more than the MCAR and minimises the number of rules for 7 datasets. The classification accuracy including the number of rules regarding the REUTERS-21578 dataset are illustrated in Figures 5 and 6.

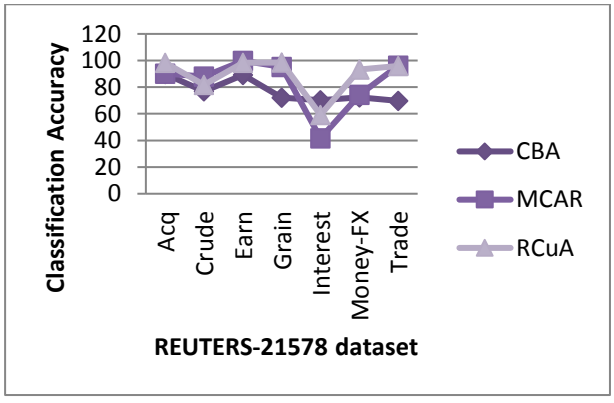


Fig. 5: The Classification Accuracy of the REUTERS-21578 Dataset.

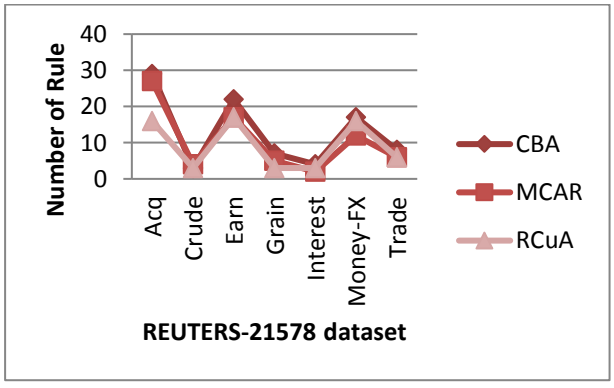


Fig. 6: The Number of Rules Regarding the REUTERS-21578 Dataset

5. Conclusion

In this paper, a novel text classification model that depends on the Associative Classification (AC) method, namely, the Rule Classification use Association (RCuA) is produced. Since the research is a key role for a specific aim of an oriented acting, selecting an appropriate research approach and method is a significant mission for achieving the major aim. The model comprises five stages, which are: the data pre-processing, data representation, rule discover and production, development and evaluation. The research highlights the stages regarding the proposed model, including various stages for improving the method that can, in return, develop the association of the classification mining and cope with structured and unstructured data. The initial test examines the structured UCI dataset where the findings show an increased competitiveness in comparison with other existing algorithms (e.g. the MCAR and CBA algorithms) according to the estimation accuracy and the number of rules. The second test examines the unstructured data Reuters-21578 dataset where the findings show an increased competitiveness in comparison with the MCAR and CBA algorithms according to the estimation accuracy and the number of rules. Based on the findings of the structured dataset, the RCuA algorithm results with an average accuracy of 83.945% in comparison with the results incurred in the CBA and MCAR algorithms, which reach 82.34% and 83.655%, respectively. In terms of the unstructured dataset, the RCuA algorithm reaches an average accuracy of 89.328% in comparison with the average accuracy in the CBA algorithm, which reaches 77.34% and the average accuracy in the MCAR algorithm, which reaches 83.64286%.

6. Acknowledgement

Mohammad Hyael Al-Refai would like to give their special thanks to the Department of Information System & Technology at Sur University College for their timely support to this work.

References

[1] Fayyad, U., Gregory, P., and Padhraic, S. 1996. From data mining to knowledge discovery in databases. AI Magazine. 17 (3):37-54. DOI: 0738-4602-1996.

[2] Wanjiang, H., Tianbo, L., Yi, S., Ye, L., Xiao, H., Weijian, L., and Chi, L. 2013. Research on the Problem Model of GUI based on Knowledge Discovery in Database. International Conference on Software Engineering and Computer Science. Advances in Intelligent Systems Research. https://doi.org/10.2991/icsecs-13.2013.2.

[3] Tao, D., Weinan C., and Wenqian S., 2012. The Research of kNN Text Categorization Algorithm Based on Eager Learning. International Conference on Industrial Control and Electronics Engineering. 23-25 Aug. 2012. pp. 1120-1123. Xi'an, China. https://doi.org/10.1109/ICICEE.2012.297.

[4] Andreas, C., Kypros, H., Argyro, S.,Kleanthis, C.,Gianna, L.,Costas, K., and Christos, N., 2012. Artificial Neural Networks to Investigate the Importance and the Sensitivity to Various Parameters Used for the Prediction of Chromosomal Abnormalities. IFIP International Conference on Artificial Intelligence Applications and Innovations. Springer Berlin Heidelberg. Berlin, Heidelberg. PP:46—55. DOI.org/10.1007/978-3-642-33412-2_5.

[5] Ian H. W., Eibe, F. 2005. Data Mining: Practical machine learning tools and techniques. 2nd Ed. Elsevier. ISBN: 0-12-088407-0.

[6] Bharath., S., David, F., Engin, D., Hakan, F., and Murat, D., 2010. Short text classification in twitter to improve information filtering. 33rd international ACM SIGIR conference on Research and development in information retrieval. July 19 - 23, 2010. ACM New York. PP: 841-842. Geneva, Switzerland. ISBN: 978-1-4503-0153-4. https://doi.org/10.1145/1835449.1835643.

[7] Gyorgy, J. S., Vipin, K., and Peter W. Li., 2011. A simple statistical model and association rule filtering for classification. 17th ACM SIGKDD international conference on Knowledge discovery and data mining. PP: 823-831.San Diego, California, USA — August 21 - 24, 2011. ISBN: 978-1-4503-0813-7. https://doi.org/10.1145/2020408.2020550.

[8] Wen, Z., Taketoshi, Y., and Xijin, T., 2011. A comparative study of TF*IDF, LSI and multi-words for text classification. Expert Systems with Applications. 38(3). March 2011, PP: 2758-2765.

[9] Fadi, T., Omar, G., and Rashid, Z., 2012. Arabic Text Mining Using Rule Based Classification. Journal of Information \& Knowledge Management. 11(1).

[10] Svetlana, K., and Stan M., 2011. Email Classification with Co-Training. in Proceedings of the 2011 Conference of the Center for Advanced Studies on Collaborative Research, 2011, pp. 301-312.

[11] Dağ, H., Sayin, K. E., Yenidoğan, I., Albayrak, S., and Acar, C., 2012. Comparison of feature selection algorithms for medical data. 2012 International Symposium on Innovations in Intelligent Systems and Applications. 2-4 July 2012. Trabzon, Turkey. https://doi.org/10.1109/INISTA.2012.6247011.

[12] James, A., Cooper, J., Jeffery, K., and Saake, G., 2009. Research Directions in Database Architectures for the Internet of Things: A Communication of the First International Workshop on Database Architectures for the Internet of Things (DAIT 2009)". British National Conference on Databases. Dataspace: The Final Frontier. Springer Berlin Heidelberg. PP: 225-233.

[13] Yong, Z., Tieniu, T., and Yunhong, W., 2001. Font recognition based on global texture analysis. IEEE Transactions on Pattern Analysis and Machine Intelligence. 23(10).Oct 2001. IEEE Computer Society. PP: 1192–1200. https://doi.org/10.1109/34.954608.

[14] Xiaoguang, q., and Brian, D., 2009.Web page classification: Features and algorithms. ACM computing surveys (CSUR), 41(2), Article 12, February 2009. PP: 12-43. DOI 10.1145/1459352.1459357.

[15] Ho, C. W., Robert, W., kam, F, and Kui, L., 2014. Interpreting TF-IDF Term Weights as Making Relevance Decisions. ACM Transactions on Information Systems, 26(3), Article 13. PP: 13-24.

[16] Dimitris, M., and Beat, W., 1999. Extending naïve Bayes classifiers using long item sets. The fifth ACM SIGKDD international conference on Knowledge discovery and data mining. San Diego, California, USA. August 15 - 18, 1999. PP. 165-174.

[17] Lei, S., Rada, M., and Mingjun, T., 2010. Cross language text classification by model translation and semi-supervised learning. The 2010 Conference on Empirical Methods in Natural Language Processing. PP: 1057-1067. MIT, Massachusetts, USA, 9-11 October 2010. Association for Computational Linguistics.

[18] Gentle, J. E., Härdle, W.K., Mori, Y., 2012. Handbook of computational statistics: concepts and methods: Springer, 2012. Springer Handbooks of Computational Statistics. ISBN 978-3-642-21551-3.

[19] Erik, W., Jan, O. P., and Andreas, S. W., 1995. A Neural Network Approach to Topic Spotting. in Proceedings of the Fourth Annual Symposium on Document Analysis and Information Retrieval, 1995, PP: 317-332.

[20] Jing, D., Zhengkui, L., Weiguo, Y., and Mingyu, L., 2010. Scaling up the Accuracy of Bayesian Classifier Based on Frequent Itemsets by M-estimate. International Conference on Artificial Intelligence and Computational Intelligence. Springer Berlin Heidelberg. Berlin, Heidelberg. PP: 357—364. ISBN: 978-3-642-16530-6.

[21] Fadi, T., Qazafi, M., Lee, M., and Hussein, A., 2010. A New Classification Based on Association Algorithm. Journal of Information \& Knowledge Management. 9(1) PP: 55-64.

[22] Gyorgy, J. S., Vipin, K., and Peter W. Li., 2011. A simple statistical model and association rule filtering for classification. 17th ACM SIGKDD international conference on Knowledge discovery and data mining. PP: 823-831.San Diego, California, USA — August 21 - 24, 2011. Jesse, R., Bernhard, P., Geoff, H., and Eibe, F., 2011. Classifier chains for multi-label classification. Machine learning. 85(3), PP: 333-359. Springer Berlin Heidelberg. ISSN: 1573-0565. https://doi.org/10.1007/s10994-011-5256-5.

[23] Kui, Y., Wei, D., Dan, A. S., and Xindong, W. 2012. Mining emerging patterns by streaming feature selection. In Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining (KDD '12). ACM, New York, NY, USA, 60-68. https://doi.org/10.1145/2339530.2339544.

[24] Elena, B., Silvia, C., and Paolo, G. 2004. On support thresholds in associative classification. In Proceedings of the 2004 ACM symposium on Applied computing (SAC '04). ACM, New York, NY, USA, PP: 553-558. https://doi.org/10.1145/967900.968016.

[25] Fadi, T., 2005. MCAR: multi-class classification based on association rule. The 3rd ACS/IEEE International Conference on Computer Systems and Applications. 6-6 Jan. 2005. Cairo, Egypt. PP: 33 IEEE. https://doi.org/10.1109/AICCSA.2005.1387030.

[26] Zhonghua, T., and Qin L., 2007. A New Class Based Associative Classification Algorithm. IAENG International Journal of Applied Mathematics. 1998.–36: 2, IJAM. – . 136, vol. 141, 2007.

[27] Yongwook, Y., and Gary, G. L., 2008. Text Categorization Based on Boosting Association Rules. 2008 IEEE International Confer-

ence on Semantic Computing. 4-7 Aug. 2008. Santa Clara, CA, USA. Publisher: IEEE. https://doi.org/10.1109/ICSC.2008.70.
[28] Fadi, T., 2005. MCAR: multi-class classification based on association rule. The 3rd ACS/IEEE International Conference on Computer Systems and Applications. 6-6 Jan. 2005. Cairo, Egypt. PP: 33 IEEE. https://doi.org/10.1109/AICCSA.2005.1387030.
[29] Antonie, M. L., and Osmar R. Z., 2004. Mining Positive and Negative Association Rules: An Approach for Confined Rules. European Conference on Principles of Data Mining and Knowledge Discovery. Springer Berlin Heidelberg. Berlin, Heidelberg. PP: 27—38. ISBN: 978-3-540-30116-5.
[30] Wenmin, L., Jiawei, H., and Jian, P., CMAR: accurate and efficient classification based on multiple class-association rules. Proceedings 2001 IEEE International Conference on Data Mining, San Jose, CA, USA, 2001, pp. 369-376. https://doi.org/10.1109/ICDM.2001.989541.
[31] Xiaoxin, Y., and Jiawei, H., 2003. CPAR: Classification based on predictive association rules. SIAM International Conference on Data Mining. PP: 331.
[32] Gourab, K., Monirul, I., Sirajum, M., and Faizul, B., 2008. ACN: An Associative Classifier with Negative Rules. 11th IEEE International Conference on Computational Science and Engineering. 16-18 July 2008. Sao Paulo, Brazil. https://doi.org/10.1109/CSE.2008.48.
[33] Fadi T., Peter I. C., and Yonghong, P., 2004. MMAC: a new multi-class, multi-label associative classification approach, Fourth IEEE International Conference on Data Mining (ICDM'04), Brighton, UK, 2004, pp. 217-224. https://doi.org/10.1109/ICDM.2004.10117.
[34] Fabrizio, S., 2003. Machine Learning in Automated Text Categorization. ACM Computer Survey (CSUR), 34(1), pp. 1–47.
[35] Guozhu, D., Xiuzhen, Z., Limsoon, W., and Jinyan, L., 1999. CAEP: Classification by aggregating emerging patterns. DS'99, LNAI 1721, pp. 30–42, 1999. Springer-Verlag Berlin Heidelberg. Berlin, Heidelberg. ISBN: 978-3-540-46846-2.
[36] Merz, C., and Murphy, P. 1996. UCI repository of machine learning databases. FTP from ics. uci. edu in the directory pub/machine-learning-databases.
[37] Mohamed .R, and Yuhanis, Y. 2014. Partial rule match for filtering rules in associative classification. Journal of Computer Science. 10(4). PP.570-577. doi: 10. 3844 /jcssp. 2014. 570 .577.
[38] Yuhanis, Y., and Mohamed .R. 2012. MMCAR: Modified multi-class classification based on association rule. IEEE International Conference on Information Retrieval \& Knowledge Management. 13-15 March 2012. Kuala Lumpur, Malaysia. Bing, L., Wynne, H., Yiming, M., 1998. Integrating classification and association rule mining. Knowledge discovery and data mining, American Association for Artificial Intelligence. pp. 80–86.
[39] Rakesh, A., and Ramakrishnan, S. 1994. Fast Algorithms for Mining Association Rules in Large Databases. In Proceedings of the 20th International Conference on Very Large Data Bases (VLDB '94), Jorge B. Bocca, Matthias Jarke, and Carlo Zaniolo (Eds.). Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, PP: 487-499. ISBN: 1-55860-153-8.
[40] Lewis, D., 2004. Reuters-21578. Available: http://www.daviddlewis.com/ resources/ testcollections/reuters21578/.
[41] Man, L., Chew, L. T., Jian, S., and Yue, L., 2009. Supervised and Traditional Term Weighting Methods for Automatic Text Categorization, in IEEE Transactions on Pattern Analysis and Machine Intelligence, 31(4). PP: 721-735, April 2009.
[42] Weiss, S.M., Indurkhya, N., Zhang, T., Damerau, F. 2005. Text Mining. Predictive Methods for Analyzing Unstructured Information. Springer-Verlag New York Inc. ISBN 978-0-387-34555-0. https://doi.org/10.1007/978-0-387-34555-0.
[43] Siti, S. K., Yuhanis, Y., Husniza, H., Mohammad, H. R. 2016. Text Classification Using Modified Multi Class Association Rule. JURNAL TEKNOLOGI 78.8-2 (2016). PP: 163-170.
[44] Zhun, Z., Bingru, Y., and Wei, H. 2010. Association classification algorithm based on structure sequence in protein secondary structure prediction. Expert Systems with Applications. 37(9) (September 2010), PP: 6381-6389.
[45] Fadi, T., and Suhel, H, 2013. MR-ARM: A Map-Reduce Association Rule Mining Framework. Parallel Processing Letters. 23(3), 1350012(2013).
[46] Robert, E., Schapire, Y. S., and Amit, S. 1998. Boosting and Rocchio applied to text filtering. In Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR '98). ACM, New York, NY, USA, 215-223. https://doi.org/10.1145/290941.290996.