Jurnal Teknologi

Full Paper

TEXT CLASSIFICATION USING MODIFIED MULTI CLASS ASSOCIATION RULE

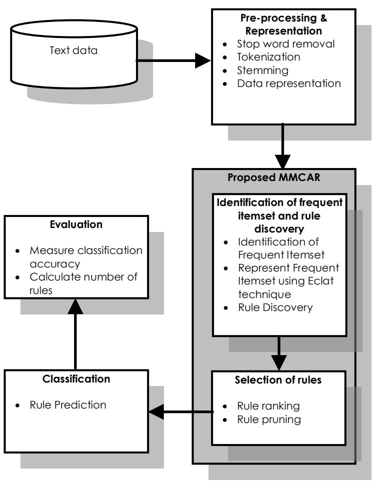
Siti Sakira Kamaruddin*, Yuhanis Yusof, Husniza Husni and Mohammad Hayel Al Refai

School of Computing, Universiti Utara Malaysia, 06010 Sintok, Kedah, Malaysia

Article history
Received
1 December 2015
Received in revised form
2 March 2016
Accepted
28 February 2016

*Corresponding author
sakira@uum.edu.my

Graphical abstract



Abstract

This paper presents text classification using a modified Multi Class Association Rule Method. The method is based on Associative Classification which combines classification with association rule discovery. Although previous work proved that Associative Classification produces better classification accuracy compared to typical classifiers, the study on applying Associative Classification to solve text classification problem are limited due to the common problem of high dimensionality of text data and this will consequently results in exponential number of generated classification rules. To overcome this problem the modified Multi-Class Association Rule Method was enhanced in two stages. In stage one the frequent pattern are represented using a proposed vertical data format to reduce the text dimensionality problem and in stage two the generated rule was pruned using a proposed Partial Rule Match to reduce the number of generated rules. The proposed method was tested on a text classification problem and the result shows that it performed better than the existing method in terms of classification accuracy and number of generated rules.

Keywords: Text mining, Frequent Pattern Mining, Associative Classification, Multi Class Association Rule.

Abstrak

Artikel ini menerangkan pengkelasan teks menggunakan kaedah Multi Class Association Rule. Kaedah ini adalah berasakanan klasifikasi bersekutu yang menggabungkan pengkelasan dan penemuan peraturan bersekutu. Walaupun penyelidikan lepas telah membuktikan bahawa klasifikasi bersekutu menghasilkan kejituan klasifikasi yang lebih baik berbanding dengan klasifikasi yang biasa, penyelidikkan yang mengkaji kebolehan kaedah ini untuk menyelesaikan masalah klasifikasi teks adalah terhad kerana masalah biasa data teks yang tinggi dimensi dan ini akan mengakibatkan penghasilan peraturan klasifikasi yang terlalu banyak. Untuk menyelesaikan masalah ini, kaedah Multi Class Association Rule telah dipertingkatkan pada dua tahap. Pada tahap pertama, corak kerap diwakilkan dengan menggunakan format data menegak untuk mengurangkan masalah dimensi tinggi data teks dan pada tahap kedua, peraturan yang dihasilkan dicantas menggunakan Padanan Peraturan Separa untuk mengurangkan bilangnan peraturan yang terhasil. Kaedah yang dicadangkan telah diuji ke atas masalah klasifikasi teks dan dapatan kajian menunjukkan bahawa kaedah yang dicadangkan telah berjaya mengatasi kaedah sediada dari segi kejituan klasifikasi dan bilangan peraturan yang dihasilkan.

Kata kunci: Pelombongan Teks, Pelombongan Corak Kerap, Klasifikasi Associative, Peraturan Association Pelbagai Kelas

1.0 INTRODUCTION

Natural language texts are overflowing the databases of world knowledge to a greater span as opposed to structured databases. Retrieving and mining relevant information from huge quantity of text data is a non-trivial task due to the high dimensionality nature of text and the lack of formal structure in the documents. The aforementioned problems had made research on text mining and its applications became prevalent. One of the main tasks in text mining is Text Classification (TC). The objective of TC is to classify incoming textual document into groups. There are various machine learning methods to perform TC such as Decision Tree, Support Vector Machine, Naive Bayes and Neural Network. Alternatively, there is a method in TC that is known as Associative classification (AC),[1-2] which represents a field of research that combines association rules discovery with classification [3-4]. The classifiers which are based on AC approach usually produces more accurate results than the classifiers based on classic data mining approaches [5-7]. The application of AC on textual data is still in infancy and is largely unexplored due to the typical problem of unstructured text data i.e. high dimensional text features that need to be represented [8-9] and the problem associated with AC methods that produces large number of rules [10]. This problem becomes more prominent when dealing with high dimensional textual data. The AC approach uses association rule mining to discover the Classification Association Rules (CARs) [3]. A key task in association rule mining is the discovery of frequent pattern which is non-trivial task when dealing with textual data. The aforementioned problems is a hindrance towards utilizing Associative Classification to solve text classification problem since it will produce an oversized classifier and limits its use for practical application.

In this research we seek to extract relevant frequent pattern from textual documents using term frequency method to reduce the high dimensionality of text. A modified Multi-Class Association Rule (MMCAR) method is proposed in this study. The essence of MMCAR lies on the identification of association rules from the frequent pattern and rule selection. In MMCAR, the extracted frequent patterns are represented as TID-intersection using a vertical data format prior to performing the frequent item mining using Eclat algorithm [11]. The discovered rules are then selected with the proposed rule ranking and pruning method. Comparison with existing method is performed to determine the effectiveness of the proposed method in terms of classification accuracy and the produced number of rules.

This paper is divided into 6 sections; the next section presents the related work on frequent pattern mining techniques followed by previous work on associative classification methods. In Section 3, the materials and method used in this research is presented followed by experimental design in section 4. Section 5 presents the discussion on the results. The paper ends with a conclusion section 6.

2.0 RELATED WORK

2.1 Frequent Pattern Mining

Frequent pattern mining was first proposed in[12] for market basket analysis in the form of associative rule mining. There are three basic frequent item set mining methodologies: Apriori, FP-growth and Eclat, however there are various extensions of the algorithms which are proposed by various researchers in the field. Frequent item-sets and prior knowledge is the basis of Apriori. Frequent item-set is found through a series of iteration. New candidate item-set is generated from the previously found frequent item-set by scanning the training data. This is the essence of the Apriori algorithm [13] and its alternatives [14].

Apriori method requires significant processing time and memory. In this regard, a new association rule mining approach is introduced i.e. FP-growth that produces extremely concentrated tree consisting of frequent pattern called FP-tree which represents the transactional database [15]. The tree represents each database transaction. The tree consists of not more than one path and the path is represented by the number of frequent items in the transaction which is equal to the length of the path. A pattern growth method is applied when mining association rule in the built FP-tree. The method uses the length one patterns in the FP-tree. Links between the patterns are investigated for each frequent pattern. This information is stored in conditional FP-tree. Unlike Apriori, the candidate rule generation is not performed in FP-tree. However, there is one drawback of the FP-tree where the completed FP-tree might not conform to the size of main memory since it can be huge in accordance to the dimensions of the mined database.

An algorithm known as Eclat Algorithm is proposed in [16] in order to reduce the scans over an input database. This procedure only involves a single scan of the database to derive all frequent itemsets. A vertical transaction layout of the database is used in the Eclat Algorithm. A simple tid-list intersection is applied to obtain the frequent itemsets. One advantage of Eclat Algorithm is the elimination of complex data structure. In [11], a recent variation of the Eclat Algorithm is presented. Here, a different vertical layout named "Diffset" is proposed. Diffset stores only the differences in the transaction identifiers (tids) of the produced frequent itemset to the candidate itemset. This minimizes the memory size needed to store the tids. Instead of storing the complete tids of each itemset, the diffset approach only stores the difference between the class and its member itemsets. Two itemsets share the same class if they share a common prefix. In this research we propose to perform frequent pattern mining using a vertical data format as proposed in[11].

2.2 Associative Classification

Associative Classification (AC) is a well-known classification method, which is based on association rule. AC has received a considerable amount of attention from various research communities. However, AC algorithms are usually used on simple and medium sized University of California, Irvine (UCI) machine learning dataset [17]. The application of AC based classifier on text collections has yet to be explored comprehensively. Thus, one of the ultimate objectives of this study is to apply AC on large and unstructured text collections. Despite the volume of interest in other domain, very little experimentation has been conducted using it within the domain of text classification and only a small number of associative classification algorithms exist that are capable of tackling text classification problems.

The AC approach requires some user-specified parameters to generate rules using an association rule mining algorithm. Existing associative classification algorithms share a common problem that is massive number of rules discovered [4,18-20]. The massive number of generated rules is impractical as it slows up the processing thus limits the use of the generated rules in practical application such as text classification. Since the number of rules generated by the algorithm could reach several thousand[15, 19,21] and are possible to be redundant and not discriminative among the target classes, the generated rules need to be pruned using pruning procedures. The rules that are left after pruning are the interesting ones that will form a model or classifier, which later will be used to classify new data. However as the number of rule decreases unfortunately the classification accuracy decreases too.

Liu et al. [22] was the first to combine association rule algorithms with classification. In their work a two stages process where the first stage finds the frequent itemser by performing the Apriori Algorithm [13] whereas in the second stage the classifier is constructed. In their experiment, they show that their method produces results comparable to the popular classification methods such the decision tree. Besides that, the CMAR Algorithm as proposed in [2] is another example of AC method which does not rely on single rule instead it picks out and analyses the correlation among the rules which has high confidence score. A tree named CR-tree is built to store the rules which are stored in descending order in accordance to the attribute frequency in the antecedent part of the rule.

Another AC method is the Multi Class Association Rule (MCAR) proposed in [7]. This method uses a rule ranking technique where only the rules that are high in confidence are stored for use in the classification stage. There are two stages in MCAR i.e. rule production and classifier construction. The advantage of MCAR is its ability to generate rules with multiple classes from data sets where each data objects is associated with just a single class. In this study, we propose to modify the MCAR to undertake the text classification problem.

3.0 MATERIAL AND METHOD

This section presents the proposed method. The design of the research was developed based on the conducted theoretical study. The overall conducts of the research is presented in this section which includes data pre-processing and representation, identification of frequent itemset and rule discovery, selection of rules, classification and evaluation. Figure 1 shows the aforementioned stages.

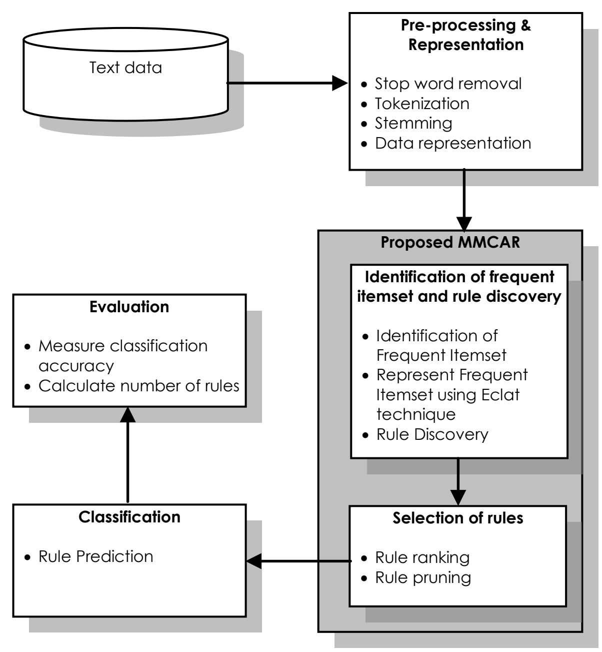


Figure 1 Research Design

The 20 Newsgroup data is used in the experiment. The dataset consists of 20,000 newsgroup documents and these documents were divided into 20 groups, based on their topic.

Table 1 shows the dataset, arranged according to its topics. There are 6 categories, i.e. computers, recreations, science, miscellaneous, politics and religion. The data is challenging as it is moderately large and sparse. Each word is treated as a single input attribute and each document contains a small fraction of the vocabulary. In addition, the labels contain noise where often the same document would fit into many newsgroups. This study presents the outcome of 1 newsgroup from each of the six categories i.e. comp.os.ms-windows.misc, misc.forsale, rec.autos, sci.med, talk.politics.misc and talk.religion.misc. Each of this dataset consists of 10,000 documents. Therefore, our data collection consists of 60,000 documents.

As shown in Figure-1, the first stage of the research design is the data pre-processing and representation. In this stage, the dataset undergoes the process of pre-processing which includes tokenization, stop word removal and stemming. Tokenization is the process of breaking the sequence of words in a sentence into individual tokens. Stop word removal is the process of removing words such as prepositions and determiners which are considered as words that are not meaningful for the learning algorithm. Stemming is the process of converting words into their root, for instance, the word "playing" is stemmed into the word "play".

Table 1 20Newsgroup Dataset

COMPUTERS	RECREATIONAL	SCIENCE
comp.graphics comp.os.ms-windows.misc comp.sys.ibm.pc.hardware comp.sys.mac.hardware comp.windows.x	rec.autos rec.motorcycles rec.sport.baseball rec.sport.hockey	sci.crypt sci.electronics sci.med sci.space
MISCELLANEOUS	**POLITICS**	**RELIGION**
misc.forsale	talk.politics.misc talk.politics.guns talk.politics.mideast	talk.religion.misc alt.atheism soc.religion.christian

The next step in this stage is data representation. A combination of vertical and horizontal layout data format is used in this work. In our work, an item is represented by the line number of which the first item occurs in the data set as well as the column number of that item. According to this representation each first appearance of a word in the text will be assigned a unique integer identification number. Therefore if we have 5 words in the text data, the unique integer identification number is as shown in Table2

Table 2 Unique Id for each word

Unique integer identification number	1	2	3	4	5
First appearance of word in text	sea	port	wind	aqaba	corn

Hence, if the indexed word appears in the text documents as shown in Table 3, the representation for the words are as shown in Table 4. The TID in Table 4 refers to the transaction ID.

Table 3 Words appearing in text

Line	Words in text (Items)
1	sea,port , wind
2	port,aqaba
3	port,corn
4	sea,port, aqaba
5	sea,corn

Table 4 ItemId for each word

TID	Item 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

The item id is represented as ColumnID(RowID) where ColumnID refers to the word's unique identification number as shown in Table 2 and RowID refers to the line number where the word first appears, for example, the words sea port and wind all appears in line 1 as shown in Table 3 whereas the first appearance of the word aqaba is in line 2 as can be seen in Table 3. Following this representation, each word in table 3 is replaced with itemID in the form of ColumID(RowID) as mentioned above.

That means, each item is converted into ColumnId, RowId representation which are simple integers and therefore the search for items to calculate the confidence values and the support values during the rule discovery process requires less complexity. This method helps to reduce the dimensionality of the textual data.

The second stage is the identification of frequent set and rule discovery. Here the first step is to identify the frequent items. In this study, the frequent items are represented using the intersection method based on the tid-list [11]. Next, the support and confidence values of ruleitems having size greater than one are computed. If the confidence value for each itemset exceeds the minimum confidence (MinConf) threshold, a single rule is generated. The generated rule is denoted as: X→ C, where C is the class with highest frequency for itemset X. These can be referred as frequent one-item. The frequent one-items are then stored in a vertical format. However, the items will be removed if they did not pass the minimum support value (MinSupp). Next, the candidate two-item is produced by using the tid-lists of the frequent one-item. This can be done by intersecting the tid-lists of any two disjoint one-items. This process is repeated in order to generate the frequent three items and subsequent frequent items. Using this method, only rules that can be represented statistically and have large confidence values aregenerated. Further details regarding this method can be referred to in [23].

This third stage is the selection of rules which involves the process of rule ranking and rule pruning. Rule ranking helps to choose the most effective rules for classification. The rules need to be sorted to give the higher quality rule a better priority, which will help to build up the classifier prior to the process of pruning the redundant rules or rules which has less confidence value. For example if a choice between two and more rules in the rule evaluation step occurs, the rule with the highest rank is selected. This means, specific rules that is ranked lower than the general rules, will never be chosen and thus removed since the general rules have included all the necessary items. Generally, rule ranking in AC is based on support, confidence and cardinality of the rule's antecedent. The next step is the rule pruning which cuts down unnecessary rules that may lead to incorrect classification [19]. This step usually happens once all rules are discovered and sorted where a procedure or more are called to prune the redundant rules. For each rule that is sorted, the proposed rule pruning algorithm evaluates the applicability of the rule against the training case. If it can match not less than one training case, the rule will get inserted into the classifier. In this method even if the rule is not fully matched (i.e. partially matched) it will not influence the results of the classifiers. Hence we proposed a rule pruning method based on partially rule match (PRM). For each training data, PRM finds the first rule that satisfies the training example by having the antecedent of the rule inside the training example. When the rule is found, the algorithm marks it and deletes the training example. PRM ignores the class similarity as it aims to reduce over learning. By doing this, PRM rule pruning method minimizes the number of training examples that will be used to make the default rule. Details of PRM can be referred in [24].

In the fourth stage the produced rules are used for text classification. The proposed classification method takes into consideration two main thresholds associated with a rule (rule confidence and rule support) as a means to distinguish a group of rules that are applicable to the test data case. In this method we use group rule classification that are divided into groups for each class; and then we calculated the average support values as well as the averages confidence values for each group. We argue that, there is more than one rule contributing to the classification. This method ensures a large number of rules are used during the matching process and therefore the class assignment decision is based on multiple rules rather than single rule as in the Multi-class Classification Association Rule (MCAR) algorithm [7]. This method has more potential in producing better classification result compared to algorithms that use a single rule.

The final stage is the evaluation where experiments on the 6 newsgroup data sets were conducted using performance measurements such as classification accuracy and produced number of rules. The effect of accuracy on the number of rules is examined. The result produced using Modified Multi-Class Association Rule (MMCAR) is compared with MCAR. The next section discusses the results.

4.0 EXPERIMENTAL DESIGN

Figure 2 shows the experimental design for this study. The text dataset undergoes pre-processing and representation as discussed in the previous section. The represented dataset were then used in the proposed MMCAR and the compared method, MCAR where the rule discovery process takes place. In MMCAR the proposed PRM is performed to prune and rank rules.

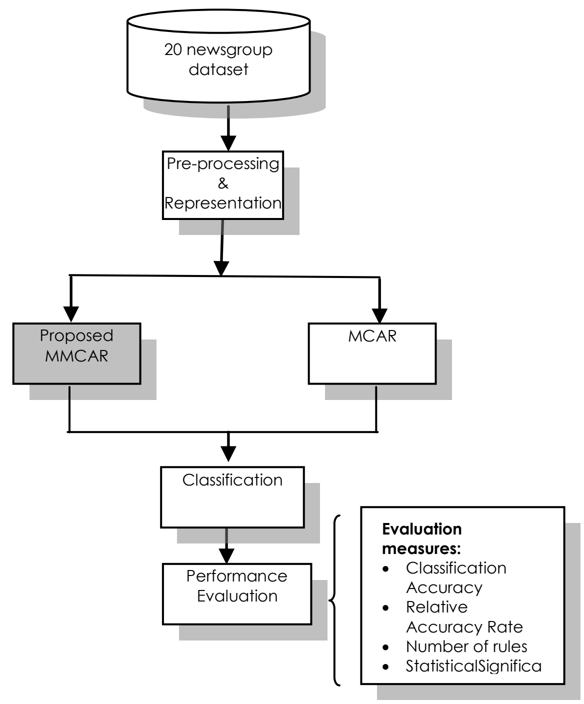


Figure 2 Experimental Design

After the execution of the proposed MMCAR and MCAR, the produced rules are used for classification task. The produced classification results are then analyzed to measure the performance. For the performance evaluation, the measurement criteria that were used are classification accuracy, relative accuracy rate, number of generated rules and statistical significance test.

The statistical significant test aims to measure the probability that the result produced by the experiment have not occurred by chance. This is done to measure the confidence of the obtained results to reflect whether the modified method produces significant result compared to the existing method.

In the conducted experiments the accuracy scores are obtained by calculating the correctly classified text data according to the news categories and the number of rules produced by the proposed method. The accuracy scores and the number of rules were compared to the existing MCAR method using the same datasets. Hence the most suitable significant test is the t-test for two samples with unequal variance.

To perform the significance test on the aforementioned measurement criteria, we first need to identify the judgment criteria by which to judge the difference between compared methods. In this research this criteria would be the accuracy scores and the produced number of rules. Hence the calculation of the difference in the mean of the accuracy scores and the produced number of rules generated by the compared methods is recognized as the judgment criteria. Next we determine the Null hypothesis. In this experiment, two Null hypotheses were defined i.e. first: "The proposed method did not produce better accuracy score than the existing method. 2:"The proposed method produced more rules than the existing method". Once the hypotheses are defined the next step is to determine the acceptance or rejection of the Null hypothesis. In order to perform this step, the t-test for correlated samples is computed using equation 1. For hypothesis 1, if the calculated t-value for the comparison is greater than the critical value for α = 0.05 (99.5%), then the Null hypotheses will be rejected (right tailed test). For hypothesis 2, if the calculated t-value for the comparison is less than the critical value for a = 0.05 (99.5%), then the Null hypotheses will be rejected (left tailed test).

5.0 RESULTS AND DISCUSSIONS

This section presents the result of the evaluation process discussed in the previous section. The proposed MMCAR was used to classify the newsgroup dataset. The resulting classification accuracy and the number of rules were recorded. The experiments were repeated using the original MCAR. Table 5 presents the results for the classification accuracy in percentage.

Table 5 Classification Accuracy

Newsgroup	MMCAR	MCAR	Difference	RAR
comp.os.ms-windows.misc	96.8	95.6	1.2	0.01
misc.forsale	93.3	92.2	1.1	0.01
rec.autos	89.5	87.8	1.7	0.02
sci.med	90.6	84.2	6.4	0.08
talk.politics.misc	91.2	87.8	3.4	0.04
soc.religion.christian	96.8	95.6	1.2	0.01
Average	93.0	90.5	2.5	0.03

From the recorded values in Table 5, we can perceive that the accuracy of MMCAR are higher than MCAR for all the 6 groups evaluated in this experiment. The highest difference in accuracy is recorded for the sci.med group where the difference is 6.4%. This may be due to the patterns found in the sci.med group i.e. the scientific terms are easily detected by the proposed method. Table 5 also reveals that the average accuracy of MMCAR is 2.5% higher than the average accuracy of MCAR.

We have calculated the Relative Accuracy Rate (RAR) to represent the variation in the accuracy rates of the proposed MMCAR compared to the result produced by the compared method MCAR. It denotes the performance of the proposed method with reference to whether it is better or worse than the compared method. A positive value of the relative accuracy rate indicates that the proposed method is better than the compared method. Relative accuracy rate is calculated using Equation 1.

$$RAR = \frac{Accuracy_{MMCAR} - Accuracy_{MCAR}}{Accuracy_{MCAR}} \quad (1)$$

Table 5 also depicts the RAR values of the comparison between MMCAR and MCAR. As can be seen in the table for all categories of the newsgroup MMCAR produces positive RAR values. This indicates that the proposed MMCAR performed better than the MCAR. Table 6 shows the number of rules produced using MMCAR and MCAR for the 6 group dataset. In Table 6, we can clearly see that the produced number of rules for the MMCAR method is smaller than the number of rules produced using MCAR method. The average number of rule of MMCAR is 13 whereas average number of rules of MCAR is 16. From these results, we can conclude that the proposed MMCAR successfully reduced the number of rules without affecting the classification accuracy for the newsgroup dataset.

Table 6 Number of rules

Newsgroup	MMCAR	MCAR
comp.os.ms-windows.misc	13	14
misc.forsale	11	13
rec.autos	17	21
sci.med	10	11
talk.politics.misc	8	8
soc.religion.christian	18	26
Average	**13**	**16**

The simple difference in the accuracy scores and number of rules is not reliable enough to determine the degree of confidence that a method is better from another method. Studies have shown that several other factors are involved such as the sample size or the number of subject being tested and the extent of variation between the sample sizes. Statistical significance testing was performed on the data that took into account the aforementioned factors. Hence Table 7 shows the result of the statistical test results.

Table 7 t-test

Measurement Criteria	Classification Accuracy	Produced Number of Rules
t-test	1.084	-0.837
Critical value	1.833	1.860

The calculated t-value for the comparison of accuracy should be higher than the critical value and the t-value for number of rules should be lower than the critical value as discussed in the hypothesis definition in experimental design section. However from Table 7 we can observe that the differences in terms of accuracy is not that significant however the difference in terms of number of rules produces a significant result since the t-value (-0.837) is lower than the critical value of 1.86. It show that the comparison in terms of number of rules produces results which is considered to be "very significant". These differences could be due to chance less than 1 out of 1000 times (0.001%). Such a value gives a very high level of confidence that the proposed MMCAR are better than MCAR in terms of producing less number of classification rules.

6.0 CONCLUSION

In this study, we employed the tid-list intersections as proposed in [11] to represent text for text classification task. This technique uses a vertical database transaction layout, where frequent itemsets are obtained by applying simple tid-lists intersections, without the need for complex data structures. To date there are limited number of researchers who have used vertical data layout for classification except for the Multi Class Association Rule (MCAR) algorithm reported in [7]. Our proposed method differs from MCAR data representation in the way of representing each item, i.e. in our study the data structure were tailored to represent text where an item is represented by the sequence of word and the line number of which the item occurs in the data set as well as the column number of that item. We have implemented the PRM rule pruning technique proposed in [24] for textual dataset and produced significant result in terms of low number of rules and comparable classification accuracy rate. Future work in this area can be targeted on handling multi-labeled text classification problem.

Acknowledgement

The authors are greatly obliged to Universiti Utara Malaysia for providing financial assistance for the research reported in this paper under the University Grant SO Code 12579.

References

[1] Dong, G., X. Zhang, L. Wong, and J. Li. 1999. CAEP: Classification By Aggregating Emerging Patterns. *In Proceedings of the Second International Conference on Discovery Science (DS '99), SetsuoArikawa and Koichi Furukawa (Eds.). Springer-Verlag, London*, UK. 30-42.

[2] Li, W., J. Han, and J. Pei, 2001. CMAR: Accurate And Efficient Classification Based On Multiple Class-Association Rules, In Proceedings of the 2001 IEEE International Conference on Data Mining (ICDM '01), Nick Cercone, Tsau Young Lin, and Xindong Wu (Eds.). *IEEE Computer Society, Washington, DC, USA*. 369-376.

[3] Simon, G. J., V. Kumar, andP. W. Li. 2011. A Simple Statistical Model And Association Rule Filtering For Classification, In Proceedings of the 17th ACM SIGKDD *International Conference on Knowledge discovery and data mining (KDD '11). ACM, New York, NY, USA*. 823-831.

[4] Thabtah, F., Q. Mahmood, L. McCluskey, H. Abdel-Jaber, 2010. A New Classification Based on Association Algorithm. *Journal of Information & Knowledge Management*. 9: 55-64.

[5] Baralis, E., S. Chusano and P. Garza, 2004. On support Thresholds In Associative Classification, *Proceedings of the 2004 ACM Symposium on Applied Computing, (SAC '04). ACM, New York, NY, USA*. 553-558.

[6] Yoon Y. & G. G. Lee, 2008. Text Categorization Based On Boosting Association Rules, *IEEE International Conference on Semantic Computing*.136-143.

[7] Thabtah, F. P. Cowling, and Y. Peng. 2005. MCAR: Multi-Class Classification Based On Association Rule, *Proceeding of the 3rd IEEE International Conference on Computer Systems and Applications*. 33.

[8] Omurca, S. I., S.Baş, E. Ekinci, 2015. An Efficient Document Categorization Approach for Turkish Based Texts. *International Journal of Intelligent Systems and Applications in Engineering*. 3(1): 7-13.

[9] Hong, S. S., W. Lee, and M. M. Han, 2015. The Feature Selection Method based on Genetic Algorithm for Efficient of Text Clustering and Text Classification. *International Journal of Advances in Soft Computing & Its Applications*. 7(1): 22-40.

[10] Abdelhamid, N., A. Ayesh, A., and F. Thabtah, 2015. Emerging Trends in Associative Classification Data Mining. *International Journal of Electronics and Electrical Engineering*. 3(1): 50-53.

[11] Zaki M. J. and K. Gouda, 2003. Fast Vertical Mining Using Diffsets. In *Proceedings of the ninth ACM Washington D.C*. 326-335.

[12] Agrawal, R., T. Imielinski, A. Swami, 1993. Mining Association Rules Between Sets Of Items In Large Databases. *In Proceedings of the 1993 ACM-SIGMOD International Conference on Management of Data (SIGMOD'93), Washington DC*. 207–216.

[13] Agrawal, R. and R. Srikant, R. 1994. Fast Algorithms For Mining Association Rules. *In Proceedings of the 20th International Conference on Very Large Data Bases, Santiago, Chile*. 487-499.

[14] Mannila, H., H. Toivonen, A. I. Verkamo, 1994. Efficient Algorithms For Discovering Association Rules. In *Proceeding of the AAAI'94 Workshop Knowledge Discovery in Databases (KDD'94), Seattle, WA*. 181–192.

[15] Han, J. and M. Kamber, 2011. Data Mining: Concepts And Techniques. *Morgan Kaufmann Pub*.

[16] Zaki, M. J., S. Parthasarathy, M. Ogihara, and W. Li, 1997. New Algorithms For Fast Discovery Of Association Rules. In *3rd KDD Conference New York*.

[17] Lichman, M. 2013. UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

[18] Thabtah F. and S. Hammoud, 2013. Mr-Arm: A Map-Reduce Association Rule Mining Framework, Parallel Processing Letters. (23): 1350012.

[19] Thabtah, F., P. Cowling and Y. Peng 2004. MMAC: A New Multi-Class, Multi-Label Associative Classification Approach. In *Proc. Fourth IEEE Int. Conf. on Data Mining (ICDM'04), Brighton, UK.* 217–224.

[20] Niu, Q., S. X. Xia and L. Zhang, 2009. Association Classification Based on Compactness of Rules. *In Second International Workshop on Knowledge Discovery and Data Mining*. 245-247.

[21] H. Ishibuchi, I. Kuwajima, and Y. Nojima, 2007. Prescreening Of Candidate Rules Using Association Rule Mining And Pareto-Optimality In Genetic Rule Selection, *Knowledge-Based Intelligent Information and Engineering Systems (4693) Lecture Notes in Computer Science*. 509-516.

[22] Liu, B., W. Hsu, and Y. Ma, 1998. Integrating Classification And Association Rule Mining. Knowledge Discovery And Data Mining. 80–86.

[23] Yusof Y. and M. H. Refai, 2012 MMCAR: Modified Multi-Class Classification Based On Association Rule, Information Retrieval & Knowledge Management (CAMP): 6-11.

[24] Refai M. H. and Y. Yusof. 2014. Partial Rule Match for Filtering Rules in Association Classification. *Journal of Computer Science.* (10): 570-577.