DETECTION OF FREQUENT ITEMS FROM THE DATASET USING UNSTRUCTURED DATA CLASSIFICATION

M. Keerthana
Department of Computer Science and Engineering, Paavai Engineering College, India

Abstract

In this article, association classification algorithms are utilised as a novel technique adopted lately by certain researchers in text categorization. Our tests using five distinct class newspapers gathered from various news sources are capable of performing well and provide higher exactness compared to the KNN, NB, and SVM. As a future study, we suggest the implementation of similar algorithms based on association rules and comparisons with ours. In addition, researchers in this field might be interested in studying the multi-labeling feature of our classification method.

Keywords:

Association classification, frequent itemset, unstructured classification, SVM

1. INTRODUCTION

Text categorization is a traditional NLP issue in the allocation of labelling and tagging to text units like sentences, queries, paragraphs, and documents. Text classification is called categorizer categorization. It offers a broad variety of applications, like answering questions, spam identification, feeling analysis, classification of news, classification of user intent, moderation of content, etc. Text data may originate from many sources; emails, social media, tickets, insurance claims, user reviews, and customer service queries and answers, to mention just a few. Text is a source of information that is very rich. However, it is difficult and time-consuming to extract insights from text since it is unstructured [6].

In recent years, machine learning models have attracted much interest. The two-step approach is followed by most traditional machine learning models. Some hand-made aspects of the papers are taken in the first stage (or any other textual unit). In the second phase, such functions are supplied to a prediction via a classifier. Popular handmade characteristics include word bags (BoW) and extensions [2]. Other popular algorithms include Nave Bayes, SVM, hidden Markov model (HMM), gradient boosting trees, and random forests. There are many limitations to the two-stage method. In order to achieve excellent efficiency, for example, dependence on handmade properties needs laborious feature design and analysis. Moreover, the significant reliance on domain knowledge for feature creation makes it difficult for the approach to generalise to future projects. Finally, the models cannot fully exploit huge quantities of training information since features (or templates of features) are pre-defined.

While the enormous models are extremely remarkable at many NLP tasks, some researchers claim that they truly do not comprehend and that they are not resilient to many fields of mission critique [10]–[14]. There has been a rising interest recently in investigating neuro-symbolic hybrid models that are not able to carry out symbolic, interpretable, neural models with certain basic constraints, such as lack of grounding.

2. RELATED WORKS

Text classification, particularly with the advent of the Internet, has been widely researched. Most algorithms are modelled on the text bag [1]. The Naive Bayes method [3] is a basic yet powerful algorithm. Different versions of Naive Bayes have been employed in text categorization, but [4] have been shown to yield superior results based on the multinomial model. Support for the categorization of texts, Vector Machines were also effectively utilised [5]. Hierarchical categorization was explored for hierarchical text data, for example, the thematic hierarchies and the Open Directory Project. The distributional clustering of terms proved to be more efficient than feature selection in text classification to counter various methods of functional selection and was first proposed by [11] when the'soft' distribution of words was applied to the classification of nouns in accordance with their conditions. We are mostly interested in "hard clustering," because our aim is to minimise the number of features and model size, where the word cluster is unique.

The authors in [7] utilised such hard clustering for text classification and, more recently, [8] used the Information Bottleneck technique for grouping words. Both [9] and [10] use comparable agglomerative clustering methods which make a greedy move in each city and demonstrate that such clusters, without any loss in classification performance using Naive Bayes, may aggressively decrease the number of features. For Support Vector Machines [3] similar findings have been observed. The Minimum Description Length (MDL) concept has been used in the agglomerate method to determine the number of clusters to be utilised for the classification task [7]. In latent semantic indexing (LSI) and its probabilistic variant, two additional dimensionality/function reduction methods are employed. Typically, these techniques are used in unattended conditions and the LSI results in less accuracy than clustering of features.

3. PROPOSED METHOD

There are two essential basic association rules metrics: support (s) and confidence (c). These are typically two preset criteria used by users to remove rules which are considered less helpful or interesting to users. Both criteria are referred to as minimal support and minimum trust. For rule Y, the rule support is the proportion of transactions containing X and Y and the rule confidence is the proportion of transaction numbers containing X [11].

The overall approach will be as follows in the suggested algorithm. Firstly, pre-processing documents will be applied for structured format conversion. The following stage will be to identify frequent items for each class label. The CACA method follows this step [12]. Thus, in documents for each label class, we

have the characteristics (words) more effectively. Only these effective words are then used to create rules. A kind of Apriorian algorithm called Bit-Apriori [13] is used in order to locate frequent things and create rules. It first stores items in binary mode in a database and then tries to discover the rules using binary operations. Using these criteria described in Fig.1, undiscovered documents are finally categorised.

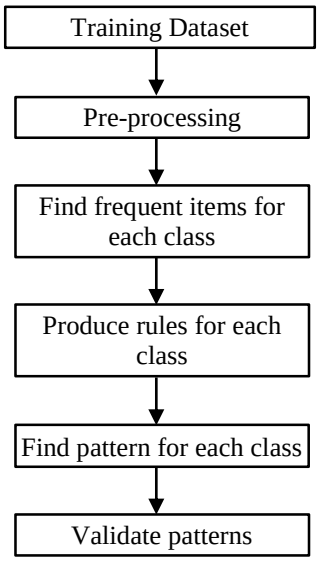

Fig.1.Steps of proposed algorithm

As stated above, the selection of features is at the core of the algorithm, which will identify the required characteristics based on words in documents and class labels.

3.1 CONVERT TEXT TO STRUCTURED FORM

Most text companies offer room for large dimensions. Since they are not relational databases, for example, for a record there is no predefined length. The average number of words in a document is much smaller than the word distribution size in text collections. By considering the words which comprise the paper as events, a document may be modelled statistically. The Bernulli Model Multivariate and the Multinomial Model [10] are two alternative models of documents. The former ignores the number of words and utilises just binary (present or not present) information, the latter incorporates the word count [10]. As it is recognised that information on numerous word occurrences does not provide substantial extra support for an accurate grading, we have combined these two approaches. This implies that the terms appear in the text first (Table.1) and then the words are ignored with fewer than the specified threshold (Table.2). The threshold is chosen on the basis that high thresholds imply fewer regulations and low thresholds imply more rules. We have established a threshold of 2 after a lot of tests, where terms used several times in a text are frequently evaluated. The chosen threshold is confirmed by a 4-fold cross validation.

Finally, we store documents in a binary table where document rows represent the most common terms used in each document. The text is tagged "1" in the table for all words in the paper. A column has been added to the database since each document's class is known to include the class value.

Table.1. Words' repetition in the documents

Document	W1	W2	W3	W4	W5	W6	Class
1	1	4	1	0	6	1	C1
2	3	3	2	1	4	5	C3
3	2	5	3	4	3	4	C3
4	5	2	1	7	1	9	C1
5	7	6	0	4	4	1	C2
6	10	3	7	9	0	0	C2
7	0	7	5	5	6	5	C1

Table.2. Binary Table containing frequent items in the documents

Document	W1	W2	W3	W4	W5	W6	C1	C2	C3
1	0	1	0	0	1	0	1	0	0
2	1	1	1	0	1	1	0	0	1
3	1	1	1	1	1	1	0	0	1
4	1	1	0	1	0	1	1	0	0
5	1	1	0	1	1	0	0	1	0
6	1	1	1	1	0	0	0	1	0
7	0	1	1	1	1	1	1	0	0

3.2 FIND FREQUENT WORDS FOR EACH CLASS LABEL

Since we aim to create rules that will predict class labels of unseen materials, for every class label we will discover effective terms. We thus seek to identify frequently used terms that have the same class designation and are repeated in numerous texts.

To accomplish this objective, the binary operation 'And' between each class column ($C[i]$) is sufficient for each word column. A count of '1's will then be generated in every column (the final row of Table.3-Table.5). In the new tables, each table cell displays common terms that are utilised and have the same class name in several texts. To make this clear, we have seven papers and three classes based on the prior example of Table.2: C1, C2 and C3, specifically. The Table.3-Table.5 show the binary "and" of each label and words following each column with the number "1." In Table.3, for instance, it can be concluded that W2 may be given in three papers; therefore, the frequency of W2 in class C1 is 3 utilising the above method.

Table.3. Frequent words in the documents for class C1

Document	C1 and W1	C1 and W2	C1 and W3	C1 and W4	C1 and W5	C1 and W6
1	0	1	0	0	1	0
2	0	0	0	0	0	0
3	0	0	0	0	0	0
4	1	1	0	1	0	1
5	0	0	0	0	0	0
6	0	0	0	0	0	0
7	0	1	1	1	1	1

Total	1	3	1	2	2	2

Table.4. Frequent words in the documents for class C2

Document	C2 and W1	C2 and W2	C2 and W3	C2 and W4	C2 and W5	C2 and W6
1	0	0	0	0	0	0
2	0	0	0	0	0	0
3	0	0	0	0	0	0
4	0	0	0	0	0	0
5	1	1	0	1	1	0
6	1	1	1	1	0	0
7	0	0	0	0	0	0
Total	2	2	1	2	1	0

Table.5. Frequent words in the documents for class C3

Document	C3 and W1	C3 and W2	C3 and W3	C3 and W4	C3 and W5	C3 and W6
1	0	0	0	0	0	0
2	1	1	1	0	1	1
3	1	1	1	1	1	1
4	0	0	0	0	0	0
5	0	0	0	0	0	0
6	0	0	0	0	0	0
7	0	0	0	0	0	0
Total	2	2	2	1	2	2

We should thus maintain a collection of phrases that predict class labels well. To achieve that, it is regarded as a frequent term for class marking if a word occurrence in the class markings is more than a threshold (Min Document Repetition) in documents linked to a class marking. In output, a document with the same class label as the class considered uses the most common term, 'True.' This step is the second selection phase. For example, when in our example we have Min Document Repetition=2:

Table.6. Frequent words for each class label

Class	W1	W2	W3	W4	W5	W6
C1	0	1	0	1	1	1
C2	1	1	0	1	0	0
C3	1	1	1	0	1	1

The suggested constant may vary depending on our requirements for Min Document Repetition. When the minimal value is evaluated, the majority of the terms will be chosen as often as possible, resulting in a wide variety of rules which are mostly inappropriate. On the contrary, a limited number of common words may be identified by taking a large constant into account that cannot fulfill our aim. The most effective results may be found in our tests by setting Min Document Repetition to 2.

3.3 SUBSETS OF FREQUENT WORDS

The Table.6 indicates the common terms that may be more helpful in forecasting the class mark of unseen texts. Thus, if we discover the rules generated by these words, we have finished the cutting step before the rules are formed. As a result, the realm of rules has decreased considerably.

It is important to notice that in our example, W2, the term often referred to as a stopword for each class label, is not included in the rule generation. Overall, common terms with class labels above 75 per cent are recognised and eliminated as "stop words."

The Apriori algorithm is being used at this level. First, each subset is created by the class labels. The rule will be formed if the generated subset (showing the common terms in the class label) is found often. A binary 'AND' operation is performed between the sub-set and the content of Table.2 to identify a subset as common terms. If it is the same as the subset, the outcome of this 'AND' operation indicates that the words in the sub-set are in an unseen document. This applies to all training documents and, ultimately, if the number of documents fulfilling this criterion exceeds the limit, the subset is considered a predictive rule and is stored for the next step.

There are two components to the created rule. The produced subset of words is put on the left side of the rule and on the right side is the name of the class label. In the generated rule, the class label in this document may be anticipated to be equal to the rule if an unknown document contains the wording of the rule.

Since all subsets of a set are created by exponential order subsets with one word and repeated by two-word subsets etc., a rule is not constructed with a certain length (for example, when no rule with the length of 5 words is produced, other subsets will not be generated). In this case, the threshold value will be reduced by one and then the regulations with the last length generated previously will be produced (after decreasing the threshold, the rules with a length of 5 will be produced again). The threshold decrease will continue until a certain constant is reached.

Now we have regulations that can forecast an invisible document's class label. We store these rules on a word table before we do nothing. A procedure identical to the previous stages is performed in order to save the created rules. In this instance, it will identify all of the words used in the rules and then it will build a table of boolean values such that its columns are terms used in the rules. The cell of a word used in the rule becomes true in every row, while the other word becomes false. This creates a table in order to make the words of the rules true in each row. In each row,

There are certain rules among the rules developed in which the component terms are the same but which predict distinct classifications. Similar rules should be removed in these situations, since the resulting outcomes are redundant and the labels of class do not forecast well.

3.4 CLASSIFYING UNSEEN DOCUMENTS

A binary tab is produced in that phase, such that its columns are the key words, the order of the key words list (words that are utilised in rules), and its rows are the indicators of unknown documents that are preprocessed. In the following phase, the cell of that word for the document becomes true when the key word is utilised on every document. It implies that in unknown texts we

will discover crucial terms. We discovered all the essential characteristics of the unknown material after filling up this table. Now all the rules (recorded in binary format) with all rows of the new table are enough to be binary 'AND.' If 'AND' is a binary value operation result, it indicates that the words of the rule are recognised as the key words of the unknown text. This algorithm may thus anticipate the document's class label.

Once this phase is complete, a series of rules are gathered which may predict the class label of the document. In the end, the projected class label will be examined by counting the rules produced for each document. The aim is to extract class labels using as many rules as possible. Generally speaking, we can find the same number of natural votes for certain classes. We generally cannot classify a news item into precise class marks in the current world. In this context, a class label is chosen randomly because of the algorithm structure, in which only one forecast should be suggested.

4. EXPERIMENTAL RESULTS

Pre-processing is the initial step of the algorithm, as in any text mining method. The primary work in the pre-processing phase helps to transform texts into structured formats. It is Tokenization, Standardization, and Stemming. Our suggested method focuses on the result of the preparation phase. Documents from any language may also be utilised to assess the performance of the algorithm since the method is binary in nature.

A dataset of Persian news items comprising 565 documents, encompassing 5 types, was gathered to illustrate the performance of the algorithm: social, financial, cultural, political, and sports. The stories were classified according to the news agency's classification.

Other data must be labelled as training data and some as test data in a controlled learning process. In order to avoid misleading outcomes in the evaluation of the techniques, these two datasets must be distinct. Consequently, many experiment runs with different datasets are typically needed at each step. The validation of an N-fold cross is one method that splits the data set and runs the experiment. Validation of the N-fold cross consists of dividing the dataset into equal N subgroups. One set is utilised at each turn to test and the remainder to train the system. The validation is 4-fold in our instance. One fold is utilised for training and learning at every turn, and three for testing so that each subset is tested once.

Preprocessing raw texts into word vectors. A Persian pre-processing method [33] is used on the dataset for this purpose. This stage will eliminate superfluous words, letters, and symbols, and will stem words with the same root. The above-mentioned methods are then utilised for document classification. The use of stemming decreased vector dimensions considerably, thereby reducing the number of rules generated and reducing algorithm time consumption thereafter.

Different ranges of minimal support and confidence will be explored for the best outcomes for BACA. The findings indicate that the optimum solution is given by taking into account the lowest value (Conf > 0) and a combination of various support ranges for distinct class labels. The reason for this outcome is because of the way we created the regulations. We first identify the effective terms and then draw up the regulations, as previously stated. Since the algorithms of association classification generate legible and comprehensible rules, domain specialists may modify them. After reviewing the rules, we see that the overall outcome is enhanced by considering different support ranges for various class labels (Table.8). It is important to highlight that the criteria utilised in our method relate to documents' words. There is a limited range of words in each text. For our suggested method, which is not linked to the quantity of categorised documents, the number of terms repeated in the documents is essential.

Table.8. Accuracy of different classes in different range of supports

Range of Supports	Financial	Social	Cultural	Political	Sport
8-6	7	62	71	56	96
7-6	56	47	68	55	96
6-6	24	84	60	80	97
6-5	24	84	60	80	97
6-4	0	91	53	86	96
Combination	61	80	52	85	97

The BACA consists of two components-initially, the creation of features, and secondly, the classification method, as stated in section 4. A separate comparison of various sections of the algorithm with certain well-known techniques of performance for each component was used to better evaluate the suggested algorithm. First, we compared the features produced using the techniques TF IDF [34] and Entropy [35] with the suggested approach (BACA). It is evident that BACA performance in terms of function creation has improved in each class. BAACA showed substantial improvements in the generation of 17.51%, 15.41% and 12.83% greater characteristics for political, sports and cultural texts than Entropy. Overall, in contrast with BACA and Entropy, TF IDF does not have high accuracy for all classes.

The SVM, NB and KNN algorithms [5] have been utilised to assess the performance of the second portion of the proposed method. Distance is cosine in KNN and weight is LTC in the KNN algorithm. Cross validation was also utilised to establish k=20 for the number of closest neighbours. These methods are used for classification in the same state as BACA after preprocessing and producing features.

5. CONCLUSION

A novel technique employed lately by certain researchers is the use of association classification algorithms in text categorization. BACA can perform well, and shows more precision than KNN, NB, and SVM in our tests using a dataset of texts in five distinct classifications, gathered from a variety of news agencies. We suggest in the future that algorithms based on similar rules of association should be implemented and their performance should be comparable to ours. Furthermore, a researcher in this field might be interested in exploring the multi-labeling feature of our classification method.

REFERENCES

[1] S. Aljawarneh, M. Aldwairi and M.B. Yassein, "AnomalyBased Intrusion Detection System through Feature Selection Analysis and Building Hybrid Efficient

Model", *Journal of Computational Science*, Vol. 25, pp. 152-160, 2018.

[2] X. Wang, W. Huang, S. Wang, J. Zhang and C. Hu, "Delay and Capacity Tradeoff Analysis for Motioncast", *IEEE/ACM Transactions on Networking*, Vol. 19, No. 5, pp. 1354-1367, 2011.

[3] Mostaque Md. Morshedur Hassan, "Current Studies on Intrusion Detection System, Genetic Algorithm and Fuzzy Logic", *International Journal of Distributed and Parallel Systems*, Vol. 4, No. 2, pp. 35-47, 2013.

[4] Ajith Abraham, Ravi Jain, Johnson Thomas and Sang Yong Han, "D-SCIDS: Distributed Soft Computing Intrusion Detection System", *Journal of Network and Computer Applications*, Vol. 30, No. 1, pp. 81-98, 2007.

[5] E.M. Yang, H.J. Lee and C.H. Seo, "Comparison of Detection Performance of Intrusion Detection System using Fuzzy and Artificial Neural Network", *Journal of Digital Convergence*, Vol. 15, No. 6, pp. 391-398, 2017.

[6] W. Fang, M. Lu and Q. Luo, "Frequent Itemset Mining on Graphics Processors", *Proceedings of 5th International Workshop on Data Management on New Hardware*, pp. 1-7, 2009.

[7] R. Agrawal and R. Srikant, "Fast Algorithm for Mining Association Rules," *Proceedings of International Conference on Very Large Data Bases*, pp. 487-499, 1994.

[8] G.P. Shapiro and W.J. Frawley, "*Knowledge Discovery in Databases*", AAAI/MIT Press, 1991.

[9] Takahiko Shintani, "Mining Association Rules from Data with Missing Values by Database Partitioning and Merging", *Proceedings of the 5th IEEE/ACIS International Conference on Computer and Information Science and 1st IEEE/ACIS International Workshop on Component-Based Software Engineering, Software Architecture and Reuse*, pp. 193-200, 2006.

[10] R.E. Thevar and R. Krishnamoorthy, "A New Approach of Modified Transaction Reduction Algorithm for mining Frequent Itemset", *Proceedings of International Conference on Computer and Information Technology*, pp.1-6, 2008.

[11] R. Agrawal, T. Imielinski and A.N. Swami, "Mining Associations Rules between Sets of Items in Large Databases", *Proceedings of ACM SIGMOD Conference on Management of Data*, pp. 207-216, 1993.

[12] L. Brieman, J.H. Friedman, R.A. Olshen and C.J. Stone, "*Classification and Regression Trees*", CRC Press, 1984.

[13] A.M.B.R. Islam and Tae Sun Chung, "An Improved Frequent Pattern Tree Based Association Rule Mining Technique", *Proceedings of International Conference on Information Science and Applications*, pp. 1-8, 2011.

[14] S. Kotsiantis and D. Kanellopoulos, "Association Rules Mining: A Recent Overview", *GESTS International Transactions on Computer Science and Engineering*, Vol. 32, No. 1, pp.71-82, 2006.