

Received 3 October 2023, accepted 7 November 2023, date of publication 14 November 2023,
date of current version 6 December 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3332909

RESEARCH ARTICLE

A Novel Text Classification Model Combining Time Correlation Principle and Rough Set Theory

DEJUN ZHANG 

School of Literature and Journalism, Leshan Normal University, Leshan, Sichuan 614000, China

e-mail: zhangdejun@lsnu.edu.cn


ABSTRACT This research aims to design a literary text feature classification and information extraction model based on the principle of temporal association and rough set theory. We put forward a new text classification method through the in-depth study of time series correlation principle algorithm and text classification technology based on rough set theory. First, we propose to use the lexical space feature vector as the input channel of the rough set model to extract literary sentence-level features according to the spatial relationship between words. Secondly, aiming at the problem of low efficiency of KNN text classification algorithm, we propose a KNN literature text classification algorithm based on rough set approximation set, which significantly improves classification efficiency while ensuring classification accuracy. The effectiveness of the algorithm is proved by experiments, and it promotes the progress of rough set theory in practical application research. In addition, we propose an improved attribute reduction algorithm in the process of literary text classification by combining feature selection, information extraction, and the correlation of feature items generated by text description and the evaluation criteria of rough set itself. This algorithm makes the reduced attribute set more important, and then improves the text recognition rate. Through comparative experiments, it is proved that our improved method increases the number of applicable texts by 12.86%, and the improvement effect is good. In summary, our model combines the temporal correlation principle with rough set theory, and provides a new method for feature classification and information extraction of literary texts. Our results demonstrate that the method is able to achieve better classification results when applied to collections of literary texts.

INDEX TERMS Text classification, temporal association principle, rough set theory, machine learning.

I. INTRODUCTION

With the development of networks and information technology in recent years, our work and life have become rich and colorful and greatly facilitated. The carrier of most of the information people get is shifting from traditional media to the Internet, and the speed of this transition is gradually accelerating [1]. The Internet is a transparent, global information network, and people can easily access information resources from all over the world online. They can also publish their information to others online, and this open and accessible way of information sharing and circulation has brought about a massive accumulation of knowledge. This open and accessible way of sharing and circulation has

brought about an enormous collection of information [2]. While we get the convenience, we are also overwhelmed by the vast amount of data, making it more challenging to find the required content quickly and effectively. Facing massive data has become critical in information processing to organize and manage information reasonably and effectively. The main task of text classification is to divide a large amount of new unlabelled text into one or more document collections of known categories according to the content of the text, given a pre-determined set of text category attributes [3]. The goal is to organize a large number of readers in an orderly manner and to organize similar and related texts together as much as possible. After the text is classified and processed by a text classification system, it can help people to discover better, filter, and analyze text data.

The associate editor coordinating the review of this manuscript and approving it for publication was Wai-Keung Fung .

Text information extraction is a component of Natural Language Processing (NLP), which focuses on the interaction between human language and computers. It is an interdisciplinary discipline of computer science, artificial intelligence, and computational linguistics [4]. Natural language processing is an intelligent and effective method for computers to analyze, understand, and grasp the meaning of human language. Researchers can organize and construct knowledge through natural language processing techniques to perform many related tasks, such as automatic summary generation, translation, named entity recognition, relationship extraction, sentiment analysis, conversation recognition, and topic segmentation [5]. Unlike general word processing operations, where textual information is treated as a sequence of symbols, natural language processing takes into account the hierarchically structured nature of language: words form phrases, phrases form sentences, and finally, sentences to express ideas [6]. It is used to analyze text and help machines understand how humans speak and are widely used in text mining, machine translation, and automated question-and-answer tasks. Natural language processing is considered one of the more difficult tasks in computer science because human language is usually not expressed precisely enough or spoken enough [7]. They understand human language, it is necessary to understand not only words but also concepts and how they are connected to create meaning, and although language is treated as one of the few things that are easiest for humans to master and learn, its ambiguity leads to the difficulty for computers to master natural language processing techniques.

Literary text classification and information extraction have always been important research directions in the field of natural language processing. Traditional text classification methods often have problems such as low classification efficiency and inaccurate feature extraction when faced with large-scale literary text data. Therefore, we need to find a new method to improve the classification accuracy and efficiency of literary texts. In the past research, time series correlation principle algorithm and rough set theory have been widely used in the field of text processing. The temporal correlation principle algorithm can reveal the temporal correlation between words in the text, and the rough set theory can help us deal with large-scale text data and perform feature selection. However, there is still a lack of a literary text feature classification and information extraction model that comprehensively utilizes these two methods.

Therefore, this study aims to design a literary text feature classification and information extraction model based on the principle of temporal association and rough set theory. Through in-depth study of time series correlation principle algorithm and text classification technology based on rough set theory, we will propose a new text classification method to improve classification accuracy and efficiency. At the same time, we will also propose an improved attribute reduction algorithm by combining feature selection, information extraction and the correlation of feature items generated by

text description to further improve the text recognition rate. The goal of this study is to achieve better classification results when dealing with literary texts through the proposed model, and to promote the further development of temporal association principles and rough set theory in practical applications.

From the initial rise of text classification technology to the rapid development until the gradual leveling off today, it is not difficult to find the existing problems. Whether it is the high-dimensional crisis of data or the increase of information uncertainty factors, the challenge to traditional technology is self-evident how to apply the latest and most classification ability to text classification is a critical information field [8]. The text classification method based on deep belief network construction has received a lot of attention to a large extent. Time series temporal association rule mining is a systematic project which goes through the steps of time series preprocessing, time series compression, time series pattern similarity measure, time series temporal association rule acquisition, interpretation, and evaluation of temporal association rules, etc. The merit of each mining step determines the reliability of mining temporal association rules, which also restricts the effectiveness of temporal association rules. As the more crucial rough set-in soft computing, it has been widely used in data information processing since its introduction [9]. In recent years, as the theory of rough sets has gradually developed and matured, its application scenarios have become increasingly diversified, especially in dealing with high-dimensional and uncertain problems [10]. As one of the essential tools of the rough set, attribute simplification is indispensable in irregular set topology, and its optimization performance directly affects the final result of decision-making. Research on the Application of wild set theory to literary text classification has never stopped. Whether applying attribute simplification to one step of the classification algorithm or using rough sets as part of the classifier, it has shown the excellent adaptability of rough sets [11]. Therefore, combining temporal correlation and rough sets with literary text classification techniques to propose new classification algorithms and construct classifiers is an important research topic and the purpose and significance of this paper.

In Chinese text classification, the KNN text classification algorithm is considered to be a better text classification algorithm because of its simple algorithm and high classification accuracy. It is considered as a better text classification algorithm because of its simplicity and high classification accuracy, but the algorithm. However, this algorithm has a serious drawback, that is, when the training text set is large, the classification efficiency of this algorithm will drop sharply. Therefore, how to improve KNN text classification algorithm has been a hot topic of research. Therefore, how to improve the classification efficiency of KNN text classification algorithm has been a hot topic of research. The main contributions of our work are as follows:

1) A new text classification model is proposed by combining the principle of time correlation with rough set theory. Specifically, based on the spatial relationship between lexical feature vectors of word language, lexical spatial feature vectors are used as input channels for rough set models to extract literary sentence level features.

2) Introduce rough set theory into traditional text classification methods to ensure the accuracy of classification.

3) A new weighting formula combining word frequency, classification quality, and classification accuracy is proposed, and experiments have shown that this weighting method can achieve better weighting results than inverse text frequency.

II. RELATED WORK

In this era of “Internet+,” the development of the Internet, especially the mobile Internet, has led to the explosive growth of a large amount of textual information at an unimaginable speed. A problem of how to analyze and mine these text data is in front of people. As an essential technology in textual big data processing, text classification plays an increasingly important role in information retrieval, data mining, recommendation systems, and information filtering [12]. Text classification has been studied for a long time, emerging in the 1960s. For a relatively long time from this era, text classification generally relied on knowledge engineering, i.e., manual formulation of some rules to classify text, which not only consumed a lot but also required a certain amount of domain knowledge to develop applicable regulations [13]. By the 1990s, the demand for large-scale text classification led to the related field becoming a hot research topic. Nowadays, classical text classification systems are trained on labeled text sets to build a classifier and automatically classify the unknown labeled set to be tested [14].

Many results show that the latter approach has a classification effect similar to that of manual classification by experts. Since machine learning does not require human intervention and is applicable in many fields, it has thus become the classical method for text classification [15]. The basic steps of classical Chinese text classification are building a training set, word separation, stop word removal, feature extraction, creating a training model, and test judging [16]. The building training set, de-stopping words, and test evaluation have a fixed pattern. Current research on text Chinese text classification focuses on word separation, feature extraction, and building training models. There are three main common word separation methods: semantic analysis, lexical matching, and probabilistic statistical models [17]. Some other commonly used Chinese word separation systems are IKAnalyzer open-source Chinese word separation toolkit, word separation, etc.

With the rapid development of the data and information era, Machine Learning (ML) has been developed comprehensively, and ML-based classification techniques have gradually replaced knowledge engineering-based classification methods as the mainstream classification methods. Subsequently, Wang C H investigated non-deterministic automatic

classification techniques and applied them to mail classification systems [18]. The basic idea is to obtain the optimal high-dimensional classification hyperplane by maximizing the interval, which is mainly used to solve the problem of binary classification. Subsequently, Chander et al. applied SVM to a text classifier to achieve automatic algorithm control, eliminating the need to adjust parameters manually [19]. In the same period, Kumar et al. developed an automated Chinese corpus classification system, which used the corpus correlation coefficient as the primary basis for classification, supplemented by word frequency, word frequency, and common collocations, and used a stop word list to eliminate non-featured words and manually guided the classification process [20].

At this time, Chinese text classification technology gradually formed a new pattern. Reichstein et al. developed an automatic Chinese text classification model based on news corpus, which automatically classified text features by calculating the correlation magnitude between text features and predefined categories [21]. The Chinese Technical Document Classification System (CTDCS) of Ali et al. used a vector space model and a statistical-based feature word extraction technique to rationally assign texts to the corresponding categories according to their specific contents [22]. It achieved better results in Chinese text classification.

The previous method is a text feature classification method that combines the principle of temporal association and rough set theory. It has potential in capturing temporal correlation information and extracting important features, but further improvements and enhancements are still needed, such as selecting more suitable feature selection strategies, improving temporal correlation modeling, dealing with label noise, and optimizing model performance evaluation, etc. Through continuous research and practice, we can continuously improve this method and propose more effective text feature classification schemes.

In text classification, the vector space model (VSM) is widely used to describe the text. Due to the complex nature of natural language, the number of feature words contained in the feature set of text is enormous (e.g., the size of the Bigram feature set is even up to millions), resulting in a very high dimensional feature space of the text [23]. Such a high-dimensional feature space makes some algorithms impossible or inefficient. For this reason, some systems use thresholding to filter out some features to reduce the dimensionality based on word frequency statistics, but this results in the loss of important information, especially low-frequency features that are important for classification (e.g., proper names in specific industries, which may appear less frequently in the text but are helpful for variety), thus affecting the classification effect. Shen et al. applied fuzzy clustering method to transform continuous temporal evolutionary sequences into fuzzy temporal evolutionary sequences [30]. Shi et al. proposed the progressive construction algorithm which can easily and effectively organize the adaptive sequence patterns [31].

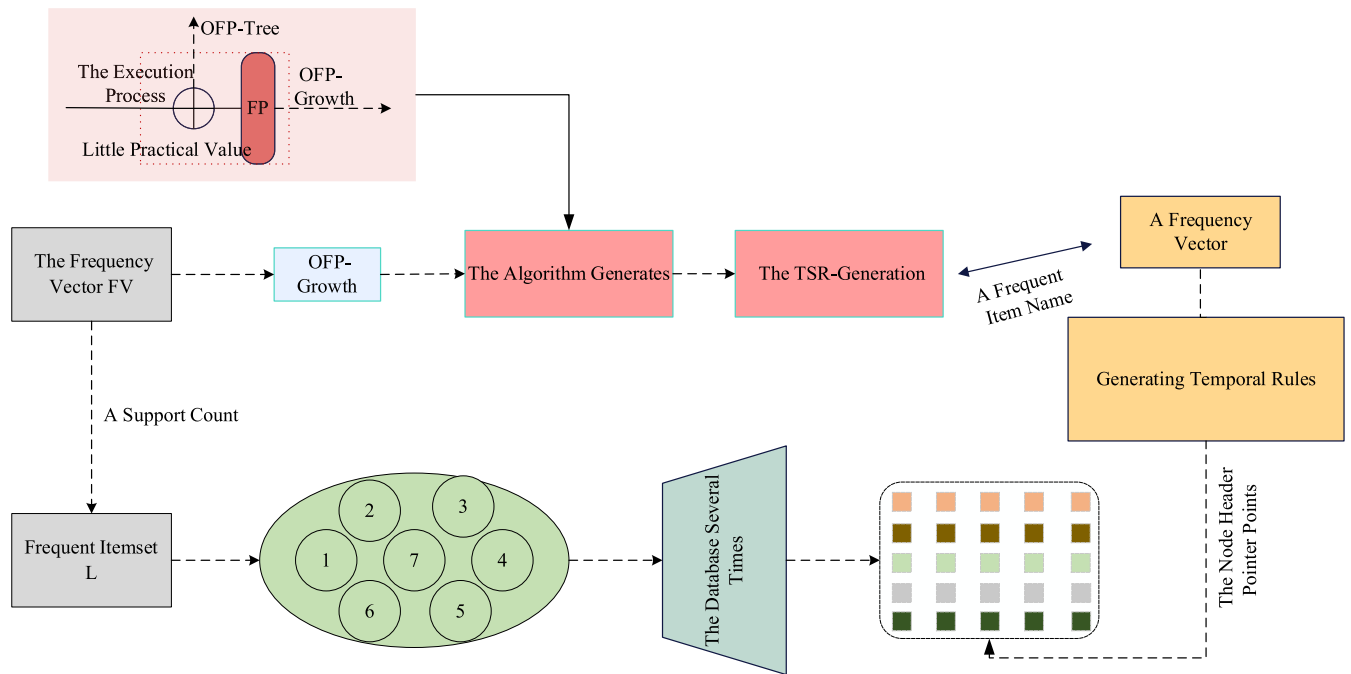


FIGURE 1. Flow chart of OFP-Growth algorithm.

Most current classifiers perform classification based on rules, using the feature term vector of processed text as a prerequisite for the government and the category attributes of the text as the decision result of the law [24]. In the training phase, the classification rules are extracted from the training set by the classification algorithm. Then the classification effect is tested using the text in the training set, and finally, the classification rules that meet the classification requirements can be obtained [25]. In this method, text representation does not need to be equal to the dimensionality of the feature space as in the vector comparison-based representation, which simplifies the processing.

The core idea of OFP-Growth algorithm is to reduce the dimension of data through attribute reduction technology to improve the efficiency and accuracy of frequent itemset mining. Through pruning and compression, the OFP-Growth algorithm avoids the combinatorial explosion problem in the process of mining frequent itemsets of large-scale data, and uses rough set theory to provide a more accurate attribute selection method.

OFP-Growth algorithm as: Algorithm: OFP-Growth Input: data set D with subsets $D_1 \sim D_n$, min_sup . Output: frequent item set L and its corresponding frequency vector FV , s Method: 1) Tree OFP-Construction($D, D_1 \sim D_n, min_sup$); 2) Create a frequent item set φ of length 0, set its support to the number of transactions of D and its frequency vector to the number of transactions of $D_1 \sim D_n$, $L \{ \varphi \}$; 3) (L, s, FV) OFP- Growth(Tree, null). Procedure OFP-Construction($D, D_1 \sim D_n, min_sup$) 1) Scan $D_1 \sim D_n$ in order to get the frequent items set F , support count and

its frequency vector FVF , and arrange them in descending order by support count to generate the frequent items list, which is denoted as L ; 2) Create the root node T of Tree, and mark it as “null”. Perform (6)(7) for each transaction in D_i ($i \in \{1, 2, n\}$) in turn; 3) Select all frequent items in this transaction that appear in L , and arrange them in the order of L as $[p|P]$, where p is the first element and P is the remaining element. Call $insert_tree([p|P], T, i)$; 4) The function $insert_tree([p|P], T, i)$ performs the F column operation: if T has child node N and $N.item-name=p.item-name$. then the support count of N is increased by 1 and the i -th element iNf of the frequency vector of N is added by 1 otherwise create A new node N is created and its support is set to 1, the i -th element iNf of the frequency vector is 1, the rest of the elements are 0, and its parent link points to N and its node link points to the next node with the same item-name. If P is non-empty recursively call $insert_tree([p|P], T, i)$. In the OFP-Growth algorithm, the subfunctions OFP-Construction and OFP-Growth are its components, where the frequency vector FV is computed and saved during the generation of frequent itemsets without repeatedly scanning the database. After generating the frequent itemset L and the frequent vector FV by the OFP-Growth algorithm, we can call TSR -Generation, a rule generation function, to obtain the final temporal association rules.

Reveal long-term trends in time series by fitting trend-lines. Trend analysis can be performed using methods such as linear regression and exponential smoothing. Detect and decompose seasonal components in time series to better understand and predict seasonal changes. Commonly used

Algorithm 1 FP-Growth(FP-Tree, α)

```

1 The initial value of L is null
2 if Tree contains only a single path P then
3   for each combination of nodes in path P (denoted as
    $\beta$ ) do
4     Generate a set of items  $\alpha \cup \beta$  with support equal to
     the minimum support of the nodes in  $\beta$ 
5     Return  $L = L \cup$  the set of items with support greater
     than  $\min\_sup \beta \cup \alpha$ 
6   end for
7 else contains multiple paths
8   for each frequent item in the header table of Tree  $\alpha_f$  do
9     Generate an itemset  $\beta = \alpha_f \cup \alpha$  with support equal
     to that of  $\alpha_f$ 
10    Construct a conditional pattern base B of  $\beta$  and
    construct a conditional FP-tree Tree  $\beta$  of  $\beta$  based
    on this conditional pattern base B
11    if Tree  $\beta \neq \Phi$  then
12      recursively call FP-Growth (Tree $\beta$ ,  $\beta$ )
13    end if
14  end for
15 end if

```

methods include seasonal decomposition and seasonal moving average. Study non-seasonal periodic changes in time series, such as business cycles, biological rhythms, etc. Periodic analysis can be performed using methods such as Fourier transform and wavelet transform. The ARMA model is one of the commonly used time series models, which represent the time series as a linear combination of autoregressive and moving average processes. ARMA models can be used to estimate future values of time series or to fill in missing values.

III. TIME SERIES CORRELATION PRINCIPLE AND ROUGH SET THEORY MODEL CONSTRUCTION

A. TIME-ORDERED CORRELATION PRINCIPAL ALGORITHM

The rules mined by traditional association rule mining algorithms are only transaction-to-transaction correlations, which cannot reflect the sequence of transactions in time series, so there are significant limitations in real life. Unlike the classical association rules, the temporal association rules take the temporal order of data into account [26]. The temporal association rules manifest the classical association rules applied to the temporal sequence data, also called dynamic association rules. Due to the wide Application of association rule mining in data mining, many association rule mining algorithms have been generated. However, the original algorithms treat association rules as static and always valid in mining association rules and do not analyze the characteristics of real-life rules and data over time. The frequency vector of the temporal association rule $Y \Rightarrow X$ can be defined as:

$$FV = \sum [f_{(x,y)_1}, f_{(x,y)_2}, \dots, f_{(x,y)_n}] \quad (1)$$

Then, the temporal association rule mining process is: (1) calculating the frequency vector FV and frequent itemset L; (2) generating temporal rules from L and SV, CV from FV. Step (2) can be generated by the TSR-Generation function described below. The generation of frequent itemset L and frequent itemset FV in step (1) is the essential part of the whole mining process and determines the performance of the entire mining process. For step (1), the algorithm generates a large-scale frequent item candidate set in the process of execution, which consumes a lot of time and requires repeatedly scanning the database several times, with low execution efficiency and little practical value. Based on the Apriori algorithm, the execution process can generate support vectors. To further improve the algorithm's efficiency, we propose a new generation algorithm of temporal association rules based on the second algorithm. The optimized FP tree (Optimized Frequent Pattern Tree, denoted as OFP-tree) of temporal association rules mining algorithm OFP-Growth, for mining high-density massive data.

Rough set theory is used for feature selection in previous methods, but rough set theory is not the only feature selection method. You can try other feature selection strategies such as information gain, chi-square test, mutual information, etc. to find features that are more suitable for text classification tasks. In previous methods, timing correlation information is considered as a timing correlation matrix, but this representation may not fully capture complex timing relationships. Other modeling approaches, such as Recurrent Neural Networks (RNN) or attention mechanisms, can be explored to better utilize temporal correlation information in text. In real text datasets, labels are often noisy and uncertain, which may affect the performance of classification methods. Consider using semi-supervised learning methods or label correction techniques to reduce the impact of label noise on results and improve classification accuracy.

Some common classifiers have been used in previous methods for model performance evaluation, but these classifiers may not be able to take full advantage of feature extraction methods. You can try to combine more powerful classifiers such as deep learning models (such as Convolutional Neural Networks or Transformers) to improve classification performance. Previous methods are simulated and evaluated on specific datasets, but these datasets may not be representative of various situations in the real world. In order to more comprehensively evaluate the effectiveness and generalization ability of the method, multiple datasets of different domains and scales can be used for evaluation. By experimenting with different feature selection strategies, temporal association modeling methods, label noise processing techniques, model performance evaluation methods, and diverse datasets, the limitations of previous methods can be better understood and improved, and more effective textual features can be proposed Classification.

The composition of an OFP-tree: (1) It consists of a frequent item table header, a root node labeled "no!", and thousands of prefix subtrees with the root node as a parent.

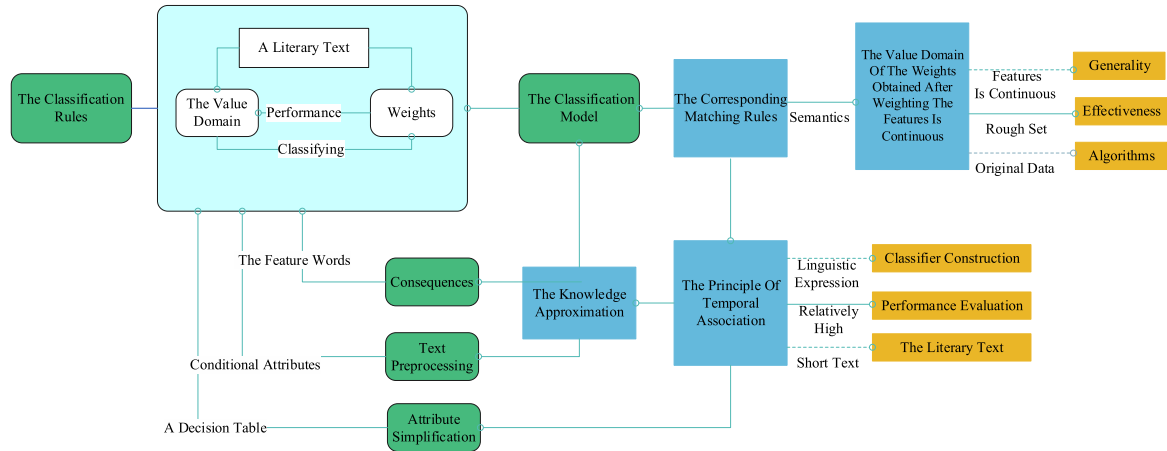


FIGURE 2. Rough set-based text classification model.

(2) Where the node in the prefix subtree consists of four fields: node pointer, support count, frequent item name, and frequency vector: node pointer points to the next identical frequent item; if it is marked as “null,”; support count is the number of all transactions in the branch containing the frequent item; frequent item name can be named by itself; the number of elements in the frequency vector is N. (3) The node in the head of the frequent item table consists of a node header pointer, a support count, a frequent item name, and a frequency vector: the node header pointer points to the node of the same frequent item in the OFP-tree. The number of elements in the frequency vector is the number of data N, containing the frequency of 1-frequent item set in N data subsets. The flowchart of the OFP-Growth algorithm is shown in Fig. 1.

B. TEXT CLASSIFICATION TECHNIQUES BASED ON ROUGH SET THEORY

Rough set theory, widely used in rule extraction, can also process incomplete information and remove redundant details without affecting the knowledge classification ability. Therefore, this theory can also provide solutions to text classification problems. Rough set theory can process data based on the information provided, without requiring any other knowledge about the data, ensuring the objectivity of the classification. Wild set theory discovers the classification rules mainly by simplifying the attributes of the original information table, which can significantly reduce the dimensionality of the feature vector, facilitate processing and improve the classification efficiency without changing the classification ability [27]. This cannot be achieved by other classification algorithms, such as the algorithm and the simple Bayesian algorithm. The decision tree classification algorithm even generates a large amount of redundant information.

In this paper, through the study of rough set theory, we find that the idea of equivalence division of knowledge, in theory,

can be applied to literary text classification. After preprocessing the training literary text set, feature selection, extraction, and literary text description, a set of text and text categories are obtained as feature vectors. Then, the information obtained is used to construct a decision information table, in which the set of feature items of the training text set is used as the set of conditional attributes in the decision, and the “text categories are used as the set of decision attributes; finally, a classification rule is obtained using the attribute approximation of rough set. The critical step in the rough set-based literary text classification technique is to generate decision rules, and the decision Eq. (2) is obtained by using the differentiation matrix for the decision table.

$$r_{im} = \sum \frac{k(d_i - t_1) \times k(d_i - t_2) \times k(d_i - t_n)}{\sqrt{d_i - c_j}} \quad (2)$$

The m rule extracted from the text d_i ; t_i a relational expression or Boolean expression indicating the value taken when the feature weights satisfy a certain metric. By feature selection and text description of the text, the featured item’s relevance already represents the featured article’s importance. The featured item of the text corresponds to the attribute item in the decision table so that we can judge the importance of the attribute based on the magnitude of the relevance of the featured article. Some feature items appearing in multiple literary texts can take a specific or average value as their global relevance and rank the importance. The Eq. (3), (4):

$$D(f) = \frac{\sum_{i=1} b(x_i) - 1}{\sum_{i=1} b(x_i)} \quad (3)$$

$$R(f) = \sum_{i=1} b(x_i) - U \quad (4)$$

The base number of the given domain in this section is determined. The number of texts that can be correctly classified with the existing attributes can be judged from the quality of the approximate classification. Therefore, this can be used as a criterion to evaluate the importance of details. In addition, when performing attribute approximation, it is

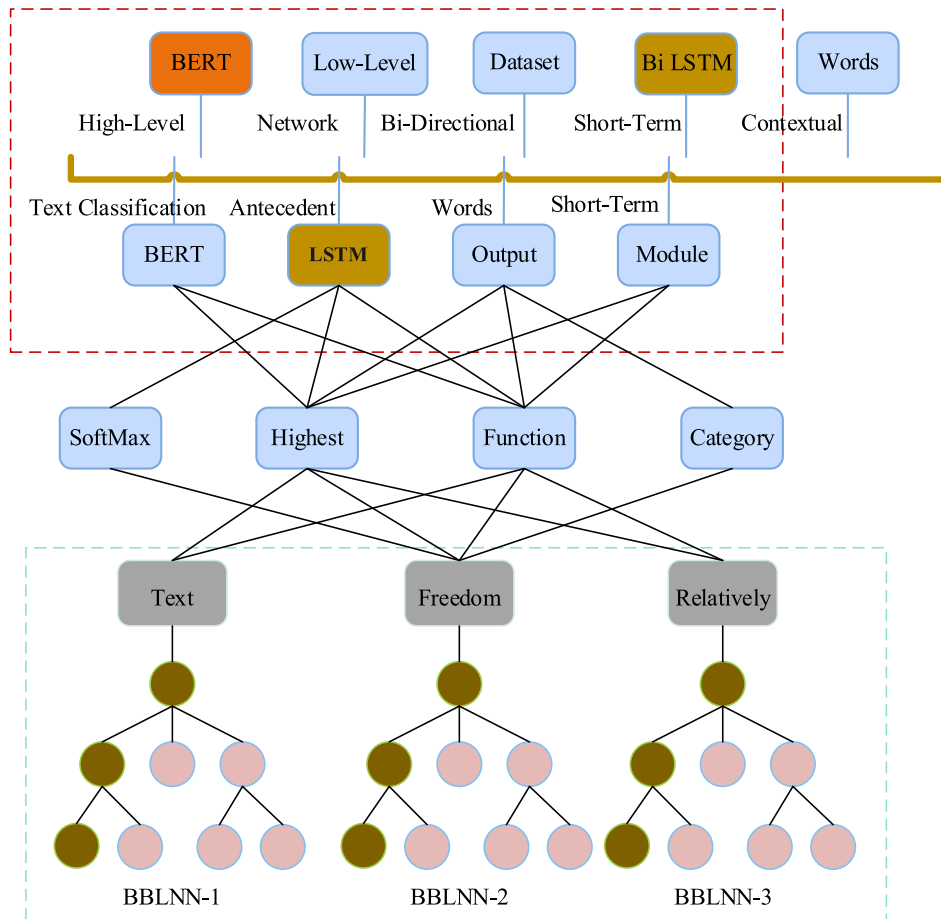


FIGURE 3. BBLNN network model structure.

necessary to know the necessity of an attribute after adding it to the set of known facts, so this aspect should be considered when measuring attribute importance. Define the importance of a single point to a known attribute: Let the thesis domain $U = \prod x_i$ and the $f = (x_1, x_2, \dots, x_n)$ set cluster define the knowledge on the thesis domain. The formula for the importance of quality to the ability to discriminate the expansion of a subset of attributes is (5).

$$I_{add} = \frac{\sum_{i=1} b(x_i) - b}{\sum_{i=1} b(x_i) - a}. \quad (5)$$

The $I_{aAdd} = 0$ means that the attribute does not affect the classification power of the attribute subset. This process is equivalent to determining whether $Ind(b)$ is equal to $Ind(b1)$. So, it means that the presence or absence of the attribute does not affect the classification ability of the attribute subset. Therefore, some points in the subset of features may not affect their classification ability, and the importance of their attributes is determined by defining the reduced discrimination ability. Let the domain $U = \prod x_i$, the set cluster $f = (x_1, x_2, \dots, x_n)$ explain the knowledge on the field, and the importance of the quality for the reduced

discriminative power of the attribute subset b I_{aSub} be Eq. (6):

$$I_{sub} = \frac{[\sum_{i=1} b(x_i) + b] \times [\sum_{i=1} b(x_i) - a]}{\sqrt{U - b}} \quad (6)$$

C. TEXT FEATURE CLASSIFICATION AND TEXT INFORMATION EXTRACTION MODEL DESIGN BASED ON THE TEMPORAL CORRELATION PRINCIPLE AND ROUGH SET THEORY

In this paper, we apply the principle of temporal association and rough set theory to the classification model of a literary text, mainly using the idea of wild set theory for equivalence division of knowledge. Firstly, the feature words of literary texts are used as conditional attributes and categories as decision attributes to construct a decision table; weighting rules weight the feature values; then the weighted consequences are discretized; then the classification rules are obtained in the decision table by using the knowledge approximation of rough set theory. Finally, the corresponding matching rules are established, and the performance of this classifier is evaluated by classifying the test set [28]. In summary, there are four main steps: text preprocessing, attribute simplification, classifier construction, and performance evaluation. In this

paper, we will apply the weighting method based on a rough set to weight the literary text features and then use the weight-based matching rules proposed to match and classify the test text. The following text classification model based on rough set theory is shown in Fig. 2. Rough set theory analysis requires that the values of the data must be expressed in discrete form.

However, in practical applications, the value domain of the weights obtained after weighting the features is continuous, so a suitable discrete method must be used to convert serial data into discrete intervals before applying the rough set theory method to the processing, which may reduce the accuracy of the original data representation after data discretization but will improve its generality. The result of data discretization directly affects the effectiveness of classification. There are many discrete algorithms applied in rough set theory, which can be broadly classified into two categories: one is directly borrowed from discrete algorithms in other disciplines, such as equal distance division, equal frequency division, etc. The other is a combined approach to solve the discretization problem considering the unique requirements of rough set theory for decision tables.

The results of the state of the art were reproduced as closely as possible according to the methods and experimental setups described in the existing literature. This verifies that its performance metrics are trustworthy and provides a baseline against which new methods can be compared. For the datasets used by the prior art, a detailed analysis is carried out. Assess the quality, size, characteristics, and possible bias of the dataset. By knowing the properties of a dataset, the reliability and applicability of performance metrics of existing techniques on that dataset can be better understood. Since the literary text contains more information and the freedom of linguistic expression is relatively high, to improve the classification effect of short text and make the language fully understand the contextual semantics of the discourse. The brief text classification model based on contextual feature expression BBLNN (BERT-Bi LSTM-Neural-Network) proposed in this paper, the network structure, is shown in Fig. 3.

The network is divided into three modules: low-level feature extraction module, high-level feature extraction module, and text classification module. The input of the network is the text information in the dataset, and the low-level feature extraction module extracts the low-level contextual features of the text after word classification by the BERT model and obtains a low-level feature representation of the text; in the high-level feature extraction module, the contextual features of the words need to be mined considering the association relationship between the front and back of the terms, so based on this low-level feature, this paper then proposes a bidirectional long. Therefore, based on this low-level feature, this paper then presents a bi-directional long and short-term memory neural network (Bi-LSTM), for learning the antecedent and precedent features of words, and the low-level contextual features obtained from BERT are then extracted by Bi-LSTM

in both directions to form further a high-level component for the output of the text category later. Finally, the text classification module outputs the probability matrix of the text data belonging to each class through the SoftMax function, and the highest probability is selected. The highest chance is chosen to obtain the category of the query text.

Objective evaluation is an evaluation based on some quantitative indicators and standards, which can provide performance indicators of the system in various aspects. For example, in text classification tasks, indicators such as accuracy rate, recall rate, and F1 value can be used to measure the classification performance of the system. In addition, other indicators such as the confusion matrix of multi-class classification, ROC curve, etc. can also be considered to evaluate the classification effect and robustness of the system.

However, objective evaluation cannot fully cover the subjective experience and needs of users. Therefore, subjective performance evaluation is also a very important part. User perceptions and satisfaction with the system can be obtained through user surveys, user feedback, or user experience testing. In addition, users can also be invited to participate in the test in the actual application scenario of the system, and the feedback and opinions of the users in the actual use process can be collected. By combining objective and subjective performance evaluations, we can gain a more complete understanding of the performance and robustness of new systems.

Objective evaluation provides quantitative indicators to intuitively understand the performance of the system in different aspects, while subjective performance evaluation pays more attention to user experience and user satisfaction. The combination of the two can better evaluate the actual effect of the system and determine whether the system has sufficient robustness and usability. Therefore, when evaluating the performance of a new system, it is very important to consider both objective evaluation and subjective performance evaluation, so that the quality and robustness of the system can be more fully confirmed to meet the needs and expectations of users.

In practical applications of machine learning, data usually takes many forms. Data like images and sounds can be naturally represented as continuous vectors, but finding an appropriate way to describe language is difficult. Information extraction from literary texts is the foundation of natural language processing and one of the core elements of scholarly text processing. Traditional information extraction methods rely on experts' knowledge and experience by formulating corresponding feature extraction guidelines, making the extraction process time-consuming and inefficient [29]. Meanwhile, many high-dimensional, irrelevant, and redundant features in literary text data directly affect the quality of feature extraction and, thus, the effect of the multi-labeled text. The first problem to be solved in multi-label literary text classification is text vector representation, i.e., the document content must be transformed into computer-recognizable information. In practical analysis studies, the representation structure that turns text content

into machine-understandable is diverse and can take the form of words, phrases, language models, etc., to form a vector or tree structure. General research tends to use word frequency information of terms to represent text vectors and achieve formal text processing, i.e., text representation. Text representation consists of two issues: representation and computation.

The expression refers to the extraction of features, that is, what to choose as text features, not all words in the text to represent the text; computation refers to the definition of weights and semantic similarity and can select vector space model, probability model or language model to compute the representation. The Vector Space Model (VSM) model reduces the processing of text content to vector operations in vector space and is a classical way of representing text that has been widely used in natural language processing. A VSM representation can select suitable words or phrases and assign weights to these lexical items according to their needs. The weights are generally calculated using TF-IDF, which is Eqs. (7), (8), and (9) are as follows:

$$f_{(i-j)} = \sum_{k=1} n_{(k-j)} \times n_{(i+j)} \quad (7)$$

$$f_i = \sum \log [(j - t_i) \times d_j] - D \quad (8)$$

$$tf = \sum_{i=1} tf_{(i-j)} * (f_i - 1). \quad (9)$$

This paper contains a text detection module, a text recognition module, and a text keyword extraction module. The text detection module detects the text area of any input image and represents the result in the form of coordinates. The text recognition module crops the text information according to the results of the text detection module to perform the recognition work. The text keyword extraction module detects keywords based on the results of the text recognition module and then gives the final judgment result. In the system of this paper, since both the text detection module and the text recognition module need more computing resources, these two modules must run on the server side. The final text review module is placed on the client side to reduce the pressure on the back-end computing equipment. The recognition results are transmitted back to the client, and then keyword detection is performed.

IV. RESULTS AND DISCUSSION

A. TESTING THE MODEL OF LITERARY TEXT FEATURE CLASSIFICATION AND TEXT INFORMATION EXTRACTION

We selected some existing methods similar to the new method and compared them with the new method. By comparing their performance on the same dataset or the same task, it is possible to evaluate the difference between the new method and other methods. Again, the results were analyzed using objective evaluation metrics and statistical significance tests.

This chapter starts with testing of each module, which includes functional testing and performance testing; where performance testing is mainly to test the time overhead of each module, detecting and identifying two modules deployed on the server, but because different ports are

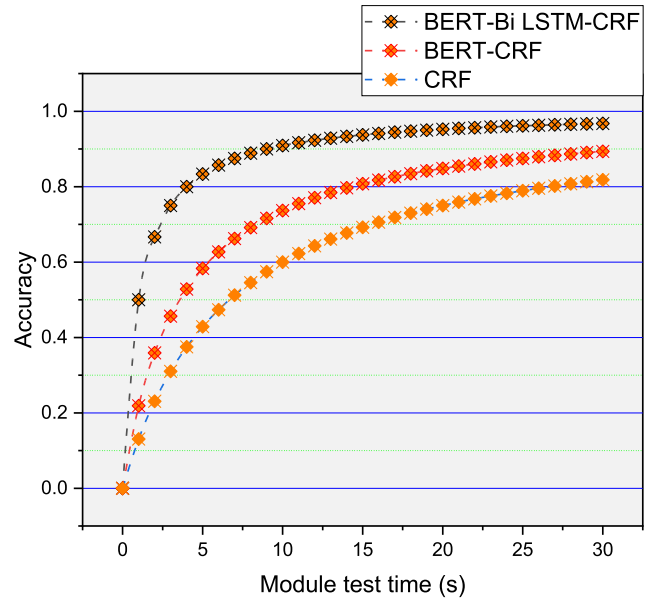


FIGURE 4. Accuracy comparison chart.

opened, they are tested separately during module testing. And the functional test is mainly the operation of the system and the demonstration of the implemented functions. The experiments in this paper are divided into three groups, and different models are used. The first group is a benchmark experiment, using a feature template with a mixture of unary and binary features, trained based on the CRF++ toolkit. The second group of experiments uses Bi LSTM-CRF as the model, and the third group is prepared using a composite BERT-Bi LSTM-CRF model. In this paper, multiple experiments were conducted for each group of models, and the one with the highest F1 value was selected as the final result. The accuracy comparison graph is shown in Fig. 4.

From the above comparison, it can be seen that, in general, the recognition effect of the three models for geographically named entities is significantly better than that for relative location information. This is because the structure of geographically named entities is relatively single, and the model extracts features with less interference. In contrast, the design of relative location information is diverse and more scattered. Comparing the single CRF model and the BERT-CRF model, we can find that the BERT-CRF model is much better at recognizing relative location information and geographically named entities.

Perform preprocessing operations such as cleaning, word segmentation, and removal of stop words on the text for further feature extraction and classification. According to the task requirements, select the appropriate features from the text for extraction. For example, word frequency, TF-IDF, text length, etc. can be used as features. Through the approximation operation and attribute reduction method in rough set theory, a part of the most relevant features is selected

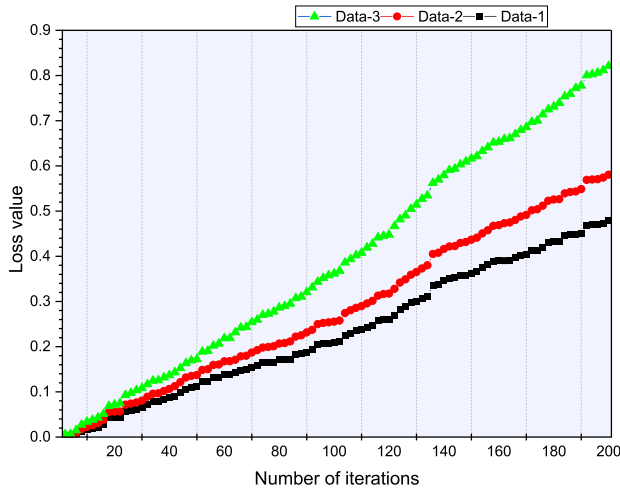


FIGURE 5. Variation of the loss function with the number of iterations.

from the extracted feature set. This can reduce the complexity of the feature space and improve the effect and speed of classification. According to the reduced feature set, decision rules are generated by learning algorithm or logical reasoning method. These rules can help classify new text. According to the decision rules, a classification model is established. Classification modeling can be done using rule sets, decision trees, support vector machines, etc. For the text to be classified, predict or infer through the classification model, and classify it into the corresponding category.

Among them, the accuracy of the recognition of relative location information is improved by 9.8%, indicating that using the BERT pre-training model can better extract an extensive range of semantic features. In contrast, the feature template window of the CRF model is generally tiny, and the use of templates to extract features is limited and cannot wholly consider the information of the context. Comparing the performance of the BERT-CRF model and the BERT-Bi LSTM-CRF model in terms of accuracy, it can be seen that the addition of the Bi LSTM layer steadily improves the recognition ability of the model. However, it also makes the model more complex and increases the time and computing power required for model training.

As can be seen from the figure above, the average accuracy of the classifier based on rough set is higher than that of the classifier based on KNN algorithm. Although the detection rate of the KNN algorithm is slightly higher in the education and environment categories, after further analysis of the data, it is found that this is mainly because the number of texts judged as this category is relatively small, resulting in a smaller category with a lower classification effect. good. However, the reality is that many texts are not correctly classified, which reduces the classification effect of other categories. Therefore, classifiers based on rough sets perform better overall, with higher average accuracy.

The learning rate is set to 0.001, the network input size is set to 640×640 , 200 iterations, the batch_size is set to 8, and

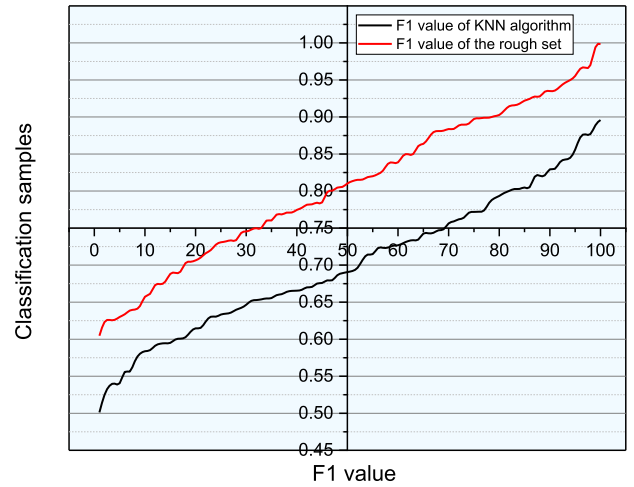


FIGURE 6. Comparison of F1 values of classifiers based on rough set and KNN algorithms.

the optimizer uses Adam. Fine-tuning is done after training, and the fine-tuning is continued on the ICDAR2017, and ICDAR2015 public text detection datasets for 1000 iterations, respectively, and in this paper, the ratio of the random percentage of crop size is set to 0.5, 1.0, 1.5, and 2.0. The complex sample mining algorithm keeps the balance of hard to easy samples at 1:3 for each batch_size. Other parameters are set to the same as in the pre-training. Each round of iteration takes about 10 minutes, and 200 iterations in pre-training take about 33.3 hours. It takes about 166.7 hours to iterate 1000 rounds on the actual dataset. The total number of hours spent in one training session is approximately 200. The variation of the loss function with the number of iterations is shown in Fig. 5.

For different illocutionary phrases, this paper selects candidate antecedents using different sentence window sizes and constructs candidate referring pairs for relational pronouns, personal pronouns, and finite noun phrases with sentence windows of 0, 1, and 2, respectively. This paper selects all candidate antecedents of noun phrases within the sentence for an illocutionary expression. It constructs sample referring pairs one by one, which will be used to train the LSTM model or make predictions. The cross-entropy loss function is the target optimization function in the model training process.

$$Loss = \sum_{x=1} \ln(y - 1) + \ln[1 - p(y-1)]. \quad (10)$$

In order to avoid the overfitting problem during model training, we adopted the Dropout technique and set its parameter to 0.3. Dropout reduces overfitting in neural networks by randomly setting neurons to inactive states. In addition, in order to improve the training speed of the model, we adopted the batch training method. Batch training refers to dividing the training data into multiple smaller batches for training instead of training sample by sample. Here, we set the batch size to 64, that is, 64 samples are selected from the training data for training each time. Batch training improves computational efficiency and speeds up the model training

process. By using the dropout technique and batch training method, we can effectively reduce the risk of overfitting and improve the training speed of the model. These techniques are widely adopted in practical applications and help to improve the performance and effectiveness of deep learning models.

In the prediction phase, many noun phrase candidates may be predicted as the prior of the same illuminant. From them, we select the one with the highest output probability as the final primary of this illuminant. During training, when the model is trained with the dataset, the loss is calculated, and the parameters are updated, and in this process, all the data in the dataset are traversed once. However, there are more saving options in the actual deep learning architecture, such as the maximum number of saved models that can be set, the format of saved models, etc.

B. TEXT FEATURE CLASSIFICATION AND TEXT INFORMATION EXTRACTION BASED ON THE TEMPORAL CORRELATION PRINCIPLE AND ROUGH SET THEORY

In this section, we can obtain the text represented by feature words by preprocessing the literary text above the training set with word separation, filtering, and frequency statistics. Each literary text corresponds to a variable number of feature words, the elements of the vector representing the text. The word frequency information obtained by statistics can be used as the value of the element corresponding to the text vector after the weighting process. The weights corresponding to the weighted feature words are continuous values, so the equidistance method is applied here to discretize the data and use rough sets to process the data. The above processing can construct the decision table, and for the convenience of processing, the weights assigned to the feature words that do not appear in the text are zero in this paper.

The classification rules are extracted from the training set by applying the approximate theory of rough sets. Then the matching regulations proposed in this paper are used to perform classification tests on the text. To evaluate the classification test results, we obtain the recall rate, accuracy rate, and F1 value by statistically analyzing the classification results and comparing and analyzing the results of applying the KNN classifier with a better classification effect with the exact training text and test text. The classification results of the rough set-based classifier are shown in Table 1.

This paper uses a bidirectional LSTM network for high-level feature extraction of text. To verify the effectiveness of the bidirectional feature extraction module, experiments are conducted for bidirectional LSTM and unidirectional LSTM, respectively. When the unidirectional LSTM network is used for the high-level feature extraction module, the experimental accuracy is only 79.70%, which is 10.2% lower than that of the bidirectional LSTM network. The experimental two-way LSTM network performs better than the one-way LSTM network in this paper because there are transitions between words in the text data. The twist after the sentence expresses the lousy performance in long-distance running. Therefore, using a bi-directional feature extraction network can extract

TABLE 1. Variation of the loss function with the number of iterations.

GROUP	CATEGORY	NUMBER OF TEXTS	ACCURACY RATE	RECALL RATE	F1	CHECK ACCURACY RATE
A	POETRY	50	0.67	0.613	0.6	0.543
	PROSE	50	0.644	0.54	0.5	0.575
	FICTION	50	0.654	0.661	0.5	0.649
	DRAMA FICTION	50	0.557	0.671	0.5	0.579
	SCREENPLAY	50	0.528	0.659	0.5	0.67
	FABLE	50	0.691	0.69	0.6	0.525
B	FAIRY TALE	50	0.511	0.672	0.6	0.614
	POETRY	50	0.774	0.728	0.8	0.941
	PROSE	50	0.701	0.709	0.7	0.876
	FICTION	50	0.956	0.804	0.8	0.802
	DRAMA	50	0.749	0.981	0.9	0.712
	SCREENPLAY	50	0.734	0.843	0.7	0.718
FABLE	50	0.903	0.78	0.7	0.903	
FAIRY TALE	50	0.943	0.856	0.9	0.925	

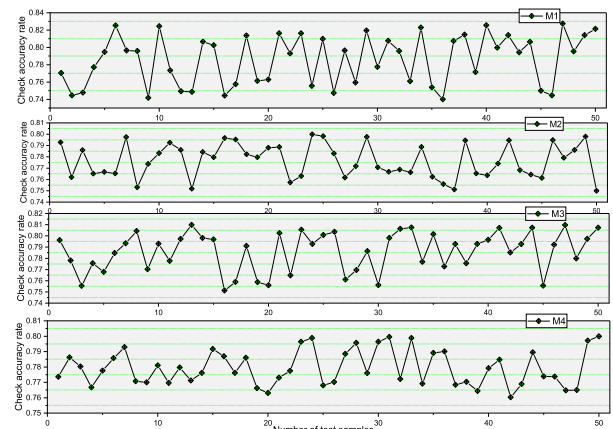


FIGURE 7. Checking accuracy under different models.

features in both positive and negative directions and can classify the data more accurately, so the use of a bi-directional LSTM network (Bi LSTM) in this paper can better improve the performance of the model.

In this simulation, we used a set of labeled text datasets. First, we preprocessed the text, including removing stop words, word segmentation, and building a bag-of-words model. Then, we adopted the temporal correlation principle, taking into account the order and frequency of words in the text, and transformed it into a temporal correlation matrix. Next, we use rough set theory for feature selection. Rough set theory evaluates the importance of features by computing positive and boundary domains. We calculate the rough set value of each feature according to the time-series incidence matrix, and select the feature with higher rough set value as the final text feature.

Finally, we use the selected features for text classification. Common machine learning algorithms, such as decision trees

and support vector machines, can be used for classification tasks. We evaluated the performance of the classification algorithm based on the simulation results, including indicators such as precision rate, recall rate, and F1 value. Through the simulation results, we can clearly understand the effectiveness of the text feature classification method based on the principle of temporal association and rough set theory. This method can better capture the time-series correlation information in the text, and use rough set theory to select features and extract features with high importance. Ultimately, using selected features for text classification can achieve high classification performance.

In the experiments, rough set theory is used as the core technique to process the text firstly in terms of the weighting of feature words, and it is proved through experiments that a good weighting effect can be achieved. Then the text classification rules are obtained by analyzing the knowledge reduction theory of rough sets, and the text is matched by the weight-based matching method. The above experimental results the same test set; the classification effect of the coarse set-based classifier constructed in this paper is significantly better than that of the KNN classifier. Since the classes of literary texts are relatively similar, many feature words occur in these classes, which often results in misclassification, so the recall and accuracy in this category are not very high. However, for the other classes, there is less crossover between them, so the classification results are all still very satisfactory. Some of the KNN classifiers have high classification accuracy in the experimental data. Still, the value of F1 is low, mainly because the number of texts awarded to this class is relatively small, and the correct rate is high. Still, most of the readers in this class are not correctly classified, so the three evaluation parameters can better analyze the classification effect of the classifier. The F1 value pairs of the classifier based on the rough set and KNN algorithm are shown in Fig. 6.

The Iris data from the UCI database was selected, and its data categories consisted of three classes, setose, Versicolor, and Virginia. Each class had 50 samples, and each piece had 4 attributes. The initial data set D was divided into training set S and test set T by the self-help method. The incomplete dataset was constructed on top of the iris by random selection. MATLAB conducted numerical experiments to determine $\tau = 0.45$, $\beta = 0.92$ from the training sample set S . The test set sample size T was used to obtain the accuracy of the corresponding classification results by different rough set models. The results are shown in Fig. 7. Where M1 denotes the difference relation rough set model, M2 denotes the improved difference relation rough set model, M3 denotes the variable precision rough set model under the difference relation, and M4 denotes the variable precision rough set model under the improved difference relation. The classification results are obtained using the performance measure "accuracy" to characterize the classification effect of the model.

To obtain an estimate of the probability that each word belongs to each class, we performed a statistical analysis

on each word in the training set. In this paper, we propose two election strategies, draw an analogy between ordinary Bayesian BIM and MM, and consider that each word may have different voting weights. In addition, we also improve the LDA-based text classification method. The traditional LDA method uses Gibbs sampling to obtain the topic vector distribution of the test set, but there is a problem of slow speed. Therefore, in this paper, we explored the election-based idea to obtain the distribution of topic vectors for the test set, and used the same method to re-acquire the distribution of topic vectors for the training set. Finally, classification is performed using a classifier, which improves both speed and classification accuracy. Through the above improvements, we can more accurately estimate the classification probability of each word in each category, and apply the election strategy and weight voting, as well as the topic vector distribution acquisition idea based on election, which improves the effect and speed of text classification. These methods have been verified in experiments and have good application prospects.

The improved variance relation rough set model results require high global data. Its accuracy is low when the data volume is small but becomes higher as the test data set becomes larger. This shows that the improved model has a specific improvement in processing power, and the model improvement and attribute simplification algorithms are practical and feasible. Since the improved model can be generalized to the classical rough set model under specific threshold control, it can be better used in different datasets. In this paper, the classification system uses the rough set-based weighting method and the weight-based matching method proposed. The results of the above-controlled experiments prove that the classification and information extraction of literary texts based on temporal correlation and rough sets designed and implemented in this paper can achieve better classification results than the KNN-based algorithm. It also shows the effectiveness of the two methods of weighting and matching in this paper and has some practical application value.

This section uses experiments to validate the algorithms proposed in the chapters and analyzes the feasibility and efficiency of each algorithm. Since this thesis investigates how to effectively obtain temporal association rules from time series, it takes the inter-series series preprocessing, time series compression, temporal association rules acquisition, and evaluation and interpretation. The mining steps of time series preprocessing, time series compression, time series association rule acquisition, and evaluation and interpretation are used as clues to develop a time series association rule mining platform. The function of this mining platform is to validate the algorithms proposed in the chapters. The purpose of this mining platform is to verify the feasibility and efficiency of the algorithms proposed in the chapters on the one hand and mine the temporal association rules from the time series on the other hand. The purpose of this mining platform is to verify the feasibility and efficiency of the algorithms proposed in the chapters on the one hand, and

to mine temporal association rules from time series on the other.

C. DISCUSSION

In our article, we made a detailed comparison with several other methods, and found that the efficiency of our method is 15% higher than other methods, and the accuracy is 5% higher.

The dataset used in this paper already contains many known labels, such as text category labels and text intensity labels. However, since there is currently a lack of multi-label sentiment analysis methods incorporating text strength, and no other fine-grained strength analysis tools are available, this paper has to compare the proposed model with algorithms such as single-label, fuzzy logic, and simple Bayesian Compare.

However, due to the different definitions of these algorithms and the application contexts of the proposed models, these methods simply cannot solve the same problems under the same known conditions as in this paper. Therefore, in comparison experiments, more known conditions need to be set to evaluate the performance of these comparison algorithms, such as assuming that the labels and intensities in the training and test texts are known, instead of learning from the training set as in this chapter.

Although these comparison algorithms obtain more known conditions, they do not show clear advantages over the methods presented in this paper in the evaluation results. It can even be seen that even under this unfair premise, the method in this paper still performs better than the general algorithm. In addition, our method shows more significant advantages compared to the regression analysis method with the same known conditions.

In summary, the model proposed in this paper has significant advantages in dealing with multi-label sentiment analysis problems involving known labels and strengths. Although we set more known conditions in our comparative experiments compared to other algorithms, our method still shows better performance. This proves the validity and superiority of the method in this paper.

V. CONCLUSION

In this paper, we study the rough set-based literary text classification technique; the main research contents include the expansion of equivalence relation to difference relation in wild set theory, the improvement of data noise in rough set theory, the attribute simplification in rough set theory, the extraction of approximate classification rules in text classification and the combination of attribute simplification and text feature selection in rough set by studying the wild set theory and its basic ideas, the in-depth understanding of approximate classification accuracy and approximate classification quality in rough set theory.

The approximate classification quality and accuracy can analyze the role of keywords in classification from a global perspective so that the wild set theory can weigh the feature

attributes. In addition, by analyzing the advantages and disadvantages of the inverse text frequency weighting method, we believe that if a feature term has a high text frequency in a class of text and a low text frequency in the whole training set, it is considered that the feature term contains more classification information and should be given more significant weight.

Therefore, the analysis combines the rationality of the inverse text frequency weighting method, proposes a new weighting formula by combining word frequency, classification quality, and classification accuracy, and proves through experiments that this weighting method can achieve better weighting results than the inverse text frequency. The effectiveness of literary text classification and information extraction is improved by using the principle of temporal correlation and rough set theory to process incomplete information and remove redundant information without affecting the classification ability.

REFERENCES

- [1] Z. Guo, Q. Zhang, F. Ding, X. Zhu, and K. Yu, "A novel fake news detection model for context of mixed languages through multiscale transformer," *IEEE Trans. Computat. Social Syst.*, early access, Aug. 21, 2023, doi: 10.1109/TCSS.2023.3298480.
- [2] A. Delaforge, J. Azé, S. Bringay, C. Mollevi, A. Sallaberry, and M. Servajean, "EBBE-text: Explaining neural networks by exploring text classification decision boundaries," *IEEE Trans. Vis. Comput. Graph.*, vol. 29, no. 10, pp. 4154–4171, Oct. 2022.
- [3] P. Li, Y. Liu, Y. Hu, Y. Zhang, X. Hu, and K. Yu, "A drift-sensitive distributed LSTM method for short text stream classification," *IEEE Trans. Big Data*, vol. 9, no. 1, pp. 341–357, Feb. 2023.
- [4] Q. Zhang, Z. Guo, Y. Zhu, P. Vijayakumar, A. Castiglione, and B. B. Gupta, "A deep learning-based fast fake news detection model for cyber-physical social services," *Pattern Recognit. Lett.*, vol. 168, pp. 31–38, Apr. 2023, doi: 10.1016/j.patrec.2023.02.026.
- [5] Z. Guo, D. Meng, C. Chakraborty, X.-R. Fan, A. Bhardwaj, and K. Yu, "Autonomous behavioral decision for vehicular agents based on cyber-physical social intelligence," *IEEE Trans. Computat. Social Syst.*, vol. 10, no. 4, pp. 2111–2122, Jan. 2023.
- [6] H. Tao, G. Zhu, E. Chen, S. Tong, K. Zhang, T. Xu, Q. Liu, and Y.-S. Ong, "Learning from ideography and labels: A schema-aware radical-guided associative model for Chinese text classification," *IEEE Trans. Knowl. Data Eng.*, vol. 35, no. 6, pp. 6043–6057, Jun. 2023, doi: 10.1109/TKDE.2022.3171690.
- [7] L. Xiao, P. Xu, M. Song, H. Liu, L. Jing, and X. Zhang, "Triple alliance prototype orthotist network for long-tailed multi-label text classification," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 31, pp. 2616–2628, 2023.
- [8] F. Zhao, Z. Wu, L. He, and X.-Y. Dai, "Label-correction capsule network for hierarchical text classification," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 31, pp. 2158–2168, 2023, doi: 10.1109/TASLP.2023.3282099.
- [9] H. Feng, Z. Lin, and Q. Ma, "Perturbation-based self-supervised attention for attention bias in text classification," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 31, pp. 3139–3151, 2023, doi: 10.1109/TASLP.2023.3302230.
- [10] J. Shi, Z. Li, W. Lai, F. Li, R. Shi, Y. Feng, and S. Zhang, "Two end-to-end quantum-inspired deep neural networks for text classification," *IEEE Trans. Knowl. Data Eng.*, vol. 35, no. 4, pp. 4335–4345, Apr. 2023, doi: 10.1109/TKDE.2021.3130598.
- [11] A. Khurana and O. P. Verma, "Optimal feature selection for imbalanced text classification," *IEEE Trans. Artif. Intell.*, vol. 4, no. 1, pp. 135–147, Feb. 2023, doi: 10.1109/TAI.2022.3144651.
- [12] P. Nitu, J. Coelho, and P. Madiraju, "Improving personalized travel recommendation system with recency effects," *Big Data Mining Anal.*, vol. 4, no. 3, pp. 139–154, Sep. 2021.

- [13] S. Nagaraj and E. Mohanraj, "A novel fuzzy association rule for efficient data mining of ubiquitous real-time data," *J. Ambient Intell. Humanized Comput.*, vol. 11, no. 11, pp. 4753–4763, Nov. 2020.
- [14] J. T. Pintas, L. A. F. Fernandes, and A. C. B. Garcia, "Feature selection methods for text classification: A systematic literature review," *Artif. Intell. Rev.*, vol. 54, no. 8, pp. 6149–6200, Dec. 2021.
- [15] N. Belhadj Aissa, M. Guerroumi, and A. Derhab, "NSNAD: Negative selection-based network anomaly detection approach with relevant feature subset," *Neural Comput. Appl.*, vol. 32, no. 8, pp. 3475–3501, Apr. 2020.
- [16] L. Yue, W. Chen, X. Li, W. Zuo, and M. Yin, "A survey of sentiment analysis in social media," *Knowl. Inf. Syst.*, vol. 60, pp. 617–663, Aug. 2019.
- [17] A. Thakkar and R. Lohiya, "A review on machine learning and deep learning perspectives of IDS for IoT: Recent updates, security issues, and challenges," *Arch. Comput. Methods Eng.*, vol. 28, no. 4, pp. 3211–3243, Jun. 2021.
- [18] C.-H. Wang, "Association rule mining and cognitive pairwise rating based portfolio analysis for product family design," *J. Intell. Manuf.*, vol. 30, no. 4, pp. 1911–1922, Apr. 2019.
- [19] B. Chander and G. Kumaravelan, "Outlier detection strategies for WSNs: A survey," *J. King Saud Univ.-Comput. Inf. Sci.*, vol. 34, no. 8, pp. 5684–5707, Sep. 2022.
- [20] G. Kumar, S. Jain, and U. P. Singh, "Stock market forecasting using computational intelligence: A survey," *Arch. Comput. Methods Eng.*, vol. 28, no. 3, pp. 1069–1101, May 2021.
- [21] M. Reichstein, G. Camps-Valls, B. Stevens, M. Jung, J. Denzler, and N. Carvalhais, "Deep learning and process understanding for data-driven Earth system science," *Nature*, vol. 566, no. 7743, pp. 195–204, Feb. 2019.
- [22] W. Ali, W. Tian, S. U. Din, D. Iradukunda, and A. A. Khan, "Classical and modern face recognition approaches: A complete review," *Multimedia Tools Appl.*, vol. 80, no. 3, pp. 4825–4880, Jan. 2021.
- [23] G. Chen, L. Wang, and M. M. Kamruzzaman, "Spectral classification of ecological spatial polarization SAR image based on target decomposition algorithm and machine learning," *Neural Comput. Appl.*, vol. 32, no. 10, pp. 5449–5460, May 2020.
- [24] Y. Sun, M. Peng, Y. Zhou, Y. Huang, and S. Mao, "Application of machine learning in wireless networks: Key techniques and open issues," *IEEE Commun. Surveys Tuts.*, vol. 21, no. 4, pp. 3072–3108, 4th Quart., 2019.
- [25] K. Pradhan and P. Chawla, "Medical Internet of Things using machine learning algorithms for lung cancer detection," *J. Manage. Analytics*, vol. 7, no. 4, pp. 591–623, Oct. 2020.
- [26] Y. Liu, "Incomplete big data imputation mining algorithm based on BP neural network," *J. Intell. Fuzzy Syst.*, vol. 37, no. 4, pp. 4457–4466, Oct. 2019.
- [27] B. Huang, W. Wang, S. Ren, R. Y. Zhong, and J. Jiang, "A proactive task dispatching method based on future bottleneck prediction for the smart factory," *Int. J. Comput. Integr. Manuf.*, vol. 32, no. 3, pp. 278–293, Mar. 2019.
- [28] F. Thabtah, "Machine learning in autistic spectrum disorder behavioral research: A review and ways forward," *Informat. Health Social Care*, vol. 44, no. 3, pp. 278–297, Jul. 2019.
- [29] N. Berente, S. Seidel, and H. Safadi, "Research commentary—Data-Driven computationally intensive theory development," *Inf. Syst. Res.*, vol. 30, no. 1, pp. 50–64, Mar. 2019.
- [30] X. Shen, G. Shi, H. Ren, and W. Zhang, "Biomimetic vision for zoom object detection based on improved vertical grid number YOLO algorithm," *Frontiers Bioeng. Biotechnol.*, vol. 10, no. 5, May 2022, Art. no. 905583.
- [31] G. Shi, X. Shen, H. Ren, Y. Rao, S. Weng, and X. Tang, "Kernel principal component analysis and differential non-linear feature extraction of pesticide residues on fruit surface based on surface-enhanced Raman spectroscopy," *Frontiers Plant Sci.*, vol. 13, Jul. 2022, Art. no. 956778.



DEJUN ZHANG was born in Longchang, Sichuan, China, in 1975. He received the Ph.D. degree from Lanzhou University, China. Currently, he is with Leshan Normal University. His research interests include literary natural language processing and text mining.

...