International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

Sequential Pattern Mining for Enhanced Multi-Label Classification

Chetna Ganesh Chand[1], Nikunj Chunilal Gamit[2], Jignesh Harenkumar Vaniya[3], Naimisha Shashikantbhai Trivedi[4], Navinkumar Ganeshan[5]**

Submitted: 06/01/2024 **Revised**: 15/02/2024 **Accepted**: 25/02/2024

Abstract: Sequential Pattern Mining (SPM) with Multi-Label Classification is a fascinating area in data mining and machine learning. It combines extracting patterns from sequential data with handling datasets where each instance is associated with multiple labels. The goal of classification is to provide the most precise prediction of an unseen instance class. A variation of single-label classification, multi-label classification uses a collection of labels linked to a single occurrence. Text classification, functional genomics, picture classification, music categorization, etc. are some examples of recent applications that use multi label classification. This article presents the subject of multi-label classification, several techniques for it, and an evaluation measures for it. Additionally, conducted comparative research of multi-label classification algorithms using both theoretical studies and simulations on different datasets.

Keywords: *Sequential Pattern Mining, Technique, Bioinformatics, Classification, Multi Label, Simulation*

1. Introduction

Sequential pattern mining is like the kid in a crowd that is trying to find a hidden pattern in everyone's behavior. It is a data mining tool that helps to spot the common sequences in the huge amounts of data. Sequential pattern is a thing that connects the events and items that are often repeated in a certain dataset in a specific order. This technique is like a jack of all trades, widely used in various areas, like for example, market basket analysis, web usage mining, text mining, and bioinformatics. [1]

Through sequential pattern mining, sales managers can reveal common purchasing patterns, thus enabling them to discern customer be haviour and build good marketing strategies. [2] This factor can be of utmost significance in the way you deal with your kids. For instance, it can reveal that customers who buy baby diapers often purchase baby food a few days later, thus enabling you to target the promotions and product placements accordingly. Through the lens of user experience, sequential pattern mining is a detective spying into and pulling out the everyday web browsing habits of users on websites. It thus allows website owners to optimize the site structure, make navigation smoother, and even provide recommendations or advertisements that are tailored to the user's specific needs and preferences. [3]Text mining letters sequential pattern mining are the applications of this, which include the extraction of frequent phrase patterns from text documents. These patterns are useful for information retrieval, text summarization, and topic modeling. [4] In bioinformatics, sequential pattern mining is the way to find the most common sub-

sequences in DNA or protein sequences. [5, 6] This is essential because, through this, scientists can get a better understanding of biological processes and can develop new drugs or treatments. [7,8] The process of sequential pattern mining typically involves several steps: Becoming so much more than just parties, lawyers, and teaching, data preprocessing, pattern generation, pattern evaluation, and pattern analysis and interpretation are now the basic tools that are enjoyed by the ones who have to take life lessons from it every day. [9, 10] A few algorithms have been created for the sequential pattern mining to be more efficient, such as GSP, Prefix Span, SPADE, and SPAM, which have made industrious effects in various fields. These algorithms utilize various techniques and fixes to deal with the large data and the complex patterns with great efficiency. [11, 12] The sequential pattern mining has really proved to be a great tool for that one to be able to find out even the most unknown and insightful stuff from the sequential data. [13, 14] This area, which is amazing and extremely useful, also confronts a lot of difficulties, such as the way it deals with noisy data, the complexity of the pattern, and the issue of scalability. Thousands of hours of work are being put into this and research is being conducted to solve these problems so that the sequential

1 Information Technology Department, Vishwakarma Government Engineering College, Ahmedabad
ORCID ID: 0009-0003-1876-8472
2 Information Technology Department, Vishwakarma Government Engineering College, Ahmedabad
ORCID ID: 0009-0004-2821-4606
3 Information Technology Department, Vishwakarma Government Engineering College, Ahmedabad
ORCID ID: 0009-0002-4121-0074
4 Information Technology Department, Vishwakarma Government Engineering College, Ahmedabad
ORCID ID: 0000-0002-8395-1586
5 Electronics & Communication Engineering, Vishwakarma Government Engineering College, Ahmedabad
ORCID ID: 0009-0002-5458-3769
***Corresponding Author Email:chetnachand87@gmail.com*

pattern mining techniques can be applied even more in different cases. [15, 16] The goal of data mining is to find valuable insights inside massive datasets. It finds previously unseen patterns in massive datasets. [17] One of the most important ideas in data mining, sequential pattern mining builds on association rule mining. It was first proposed to use sequential pattern mining [18, 19] Sequential pattern mining is to find all frequent subsequences, i.e., the subsequences whose occurrence frequency in the set of sequences is no less than min support. This is done given a set of sequences, where each sequence contains a list of elements and each element contains a set of items, and a user-specified nonsupport threshold. [20, 21] Association rule mining shows relationships between elements in the same transaction, while sequential pattern mining shows relationships between separate transactions. "Items that are regularly bought together in the same transaction may be located via association rule mining. [22, 23] In contrast, sequential pattern mining identifies products that a single consumer buys in a certain sequence across several purchases. Therefore, a marketing manager may benefit greatly from sequential pattern mining in order to determine which items a certain consumer buys in a certain order. [24]

In parallel, multi-label classification—a machine learning paradigm where each instance is associated with multiple labels—has seen growing adoption in areas such as text categorization, music genre classification, and medical diagnosis. Despite its success, applying multi-label classification to sequential datasets is challenging due to the high dimensionality and dependencies inherent in both the data and the labels.

2. Review of literature

M. J (2023) [25] the database is traversed many times. In terms of speed, the GSP method easily outperforms the Apriori algorithm. This technique does not use main memory; instead, it decides whether or not to support a candidate by scanning the dataset. It only produces candidates that fit in memory.

Minh-Thai (2022) [26] In order to extract common closed sequences, the CloFS-DBV technique is used. It compresses data, employs dynamic bit vectors, and uses a vertical data structure. Create a name for the self-pattern structure using CloFS-DBV. Pattern in CloFS-DBV For data structures, it important to think about where sequences begin and stop. Using a hybrid of DBV structure and sequence information representation, CloFS-DBVPattern achieves its structural goals. The compact data structure of the CloFS-DBV algorithm means that it requires less storage space.

3. Objectives

- To study about multi-label categorization, its techniques, and how it is evaluated.
- To analyze and empirically simulate multi-label classification algorithms on different datasets to perform a comparison analysis.

4. Statement of the problem

The majority of the real-world cases are of the instances that belong to several labels at the same time, rather than to a single class. This case, which is called multi-label classification, is different from the traditional single-label classification tasks and poses many problems for the classification. The right prediction of the labels of the instance that will be relevant to the given situation is essential in the domains like text categorization, bioinformatics, multimedia annotation, and recommendation systems. Nevertheless, the existence of label correlations, class imbalance, and high dimensionality in multi-label data give rise to significant

Difficulties to the existing classification algorithms. Besides, the process of measuring the performance of multi-label classifiers is very difficult and needs special evaluation methods that can deal with the problems of multi-label predictions. Overcoming these challenges is a must for the creation of the effective multi-label classification methods that will be capable of coping with the complexities of the real-world data and at the same time will be able to satisfy the needs of the different applications.

5. Significance of the study

The research of multi-label classification is of the highest importance nowadays when the complex relations and multiple aspects usually characterize the real-world entities. Through the research on the development and the improvement of the multi-label classification techniques application and performance, the vast opportunities of the multi-label classification techniques can be realized in various fields. Through text analysis, it can, for example, improve the topic modeling, sentiment analysis, and document categorization. Bioinformatics can be used for gene function prediction and disease risk assessment which in turn can help us to choose the best option. In multimedia applications, it can improve the image and video annotation, so that retrieval and organization of the information will be easy.

Besides, the multi-label classification can push for the development of the recommendation systems, personalized marketing, and decision support systems. Hence, this study makes a significant contribution to the creation of strong and flexible machine learning models that can precisely reflect the complex relationships in the data, thus, opening the way for the machine learning models to be used for the decision making and the solutions to be developed across the different sectors.

6. Research methodology

● Multi-Label Classification

Problems with single labels are denoted as D and L, respectively. Choose a label set l from L for every occurrence d in D. accordingly, the (d, l) single-label representation remains unchanged. The set of instances is D and the set of labels is L in multi-label issues. D, select label subset S, and L are for each-instanced. Hence, the multi-label representation: (d, S).For situations involving multiple labels, there are primarily two approaches:

Approach for transforming problems

The problem-transformation approach simplifies issues with several labels into ones with just one. In order to deal with multi-label situations; an algorithm adaptation approach expands a particular learning algorithm. An example of a multi-label issue with five class labels is shown in the table below. The given sentence is an array of lists: L= {rec, sport, swim, auto, run}.

Table 1: Example of Multi Label Problem**.**

Attributes	**A**	**A**	**A**	**A**	**B**	**B**
	B	**1**	**2**	**2**	**1**	**2**
Class Labels	**rec**	✓	✓	✓	✓	✓
	sport	✓	✓	✓	✓	
	swim	✓		✓		
	auto					✓
	run		✓	✓		

The Problem Transformation Method

The core concept of this approach is to generalize a group of single-label issues to a multi-label problem. Any classical classification algorithm can handle multi label situations since it is an algorithm independent technique. When it comes to converting multi-label issues into single-label problems, there are a number of options to choose from.

Table 2: Multi Label Example

Example	1	2	3	4
Label set	{l1, l4}	{l3, l4}	{l1}	{l2, l3,l4}

A. **Binary Relevance (BR)**

Labels are essentially classified using this way. As a result, it produces a dataset with |L| single labels from the original multi-label dataset. Each label is used to construct a binary classifier. When classifying a new instance, BR takes the positive predictions made by the L classifier and adds them together.

Table 3: Method for Binary Relevance

Ex #	l1
1	✓
2	
3	✓
4	

Ex #	l2
1	
2	
3	
4	✓

Ex #	l3
1	
2	✓
3	
4	✓

Ex #	l4
1	✓
2	✓
3	
4	✓

B. **Ranking via single label**

The multi-label dataset is converted into a single-label dataset using this procedure. For example, you may choose to reject instances with multiple labels, determine the smallest number of labels, randomly choose labels, and apply weights to each label. With each class label, a single-label classifier generates a vote (probability) that determines the ranking.

Table 4: Ranking via Single label

(a) Ignore

Ex#	Labelset
3	{l1}

(b) Maximum & Random

Maximum		Random	
Ex#	Label	Ex#	Label
1	l4	1	l4
2	l4	2	l4
3	l1	3	l1
4	l4	4	l4

(c) Copy weight

Ex#	1	1	2	2	3	4	4	4
Label	l1	l4	l3	l4	l1	l2	l3	l4
Weight	0.5	0.5	0.5	0.5	1	0.33	0.33	0.33

Ranking Via Pair-Wise Comparison (RPC)

This method compares labels in pairs. With one model for every pair of labels, it learns m=k (k-1)/2 binary models. (Where k=|L|&k is the number of labels) Model is trained relies on instances where one of the labels has been applied, but not always both. Thus, for each new instance,

we call all m models and get the ranking by tallying the votes for each label.

Table 5: one classifier for each pair of labels.

Ex#	1	3	4	
l1_l2	l1	l1	l2	
Ex#	1	2	3	4
l1_l3	l1	l3	l1	l3
Ex#	2	3	4	
l1_l4	l4	l1	l4	

Ex#	2	
l2_l3	l3	
Ex#	1	2
l2_l4	l4	l4
Ex#	1	
l3_l4	l4	

Table 6: Ranking of labels for new instance New instance x':

l1_l2	l1_l3	l1_l4	l2_l3	l2_l4	l3_l4
l1	l3	l1	l3	l2	l3

Votes for each label:

L1	l2	l3	l4
2	1	3	0

Ranking based on votes: r (l3) > r (l1)> r (l2)> r (l4)

7. Calibrated Label Ranking (CLR)

This approach builds upon the RPC technique. In order to differentiate between positive and negative labels, it adds one more virtual label V. Taking into account the votes of all labels, including virtual label V, yields the final ranking.

Table 7: Calibrated Ranking of labels.

Ex#	1	2	3	4
l1_V	l1	V	l1	V
Ex#	1	2	3	4
l2_V	V	V	V	l2
Ex#	1	2	3	4
l3_V	V	l3	V	l3
Ex#	1	2	3	4
l4_V	l4	l4	V	l4

Table 8: Ranking of labels for new instance New instance x':

l1_l2	l1_l3	l1_l4	l2_l3	l2_l4	l3_l4
l1	l1	l1	l2	l2	l4

l1_V	l2_V	l3_V	l4_V
l1	V	V	V

Votes for each label:

l1	l2	l3	l4	V
4	2	0	1	3

Ranking based on votes:

r(l1)>r (lV)>r (l2)>r (l4)>r(l3)

Label Powerset (LP)

In a multi-label dataset, this method substitutes a single label for each of the technique subsets (Distinct Label Set). New set of class labels introduced by SoL P. For example, in a multi-label dataset, the basic classifier of LP initially predicted a single label. Below Table, you can see the results of the LP approach. Labels 1, 2, and 3 and 4 are present, hence LP returns 1001.

Table 9: Label Powerset.

Ex#	1	2	3	4
Label(l1l2l3l4)	1001	11	1000	111

Pruned Set (PS)

To simplify things, this approach uses the LP method to convert multi-label datasets into single-label datasets. This multi-label dataset is pruned when the user-defined threshold, denoted as p, is applied. Those whose label sets occur less frequently than pruning parameters are pruned examples. From multi-label datasets, the PS approach extracts less significant cases. Table shows that when pruning parameter 3 is taken into account, the final row is removed.

Table 10: Pruned set method for p=3.

Label-set	l1	l2	l2, l3	l1, l4	l3, l4	l1, l2,l3
Count	16	14	12	8	7	2

Random K-Label Set (RAKEL)

The k-label technique randomly divides a big collection of labels into n smaller sets of labels, where k is the minimum size. Each label in L has an average judgment computed for training the multi-label classifier using the L-protocol. The average judgment must be greater than the threshold for the final label to be considered positive. In order to avoid LP problems, it considers the label correlation.

Table 11: Comparative Study of Problem Transformation method.

Method	Merits	Demerits
PS	Increase your speed and think about the label correlation ship.	Reliance on base classifier predictions.
CLR	It provides ranking in addition to dealing with pair-wise comparisons of each label with virtual labels.	This approach is costly theoretically. When classifying, unlabeled data is not taken into account.
RPC	Versatile approach.	Use up more memory and prediction time.
Ranking via single label	Basic in Concept.	Problems in handling label overlap.
LP	It takes label correlation ship into account.	Over fitting of training data results from this conceptually difficult strategy.
BR	Fast and simple binary classification.	Does not take label correlation into account.
RAKEL	More straightforward, takes label correlation ship into account, and has better predictive skills.	Takes longer and ignores unlabeled data when classifying.

- **Algorithm adaptation method**

This method involves developing multi-label classifiers from single-label classifiers in order to handle problems with multiple labels. Therefore, this approach relies on algorithms. Many different algorithm adaptation methods have been developed.

A. Multi-Label decision tree (C4.5)

For handling data with multiple labels, this algorithm is an extension of the basic decision tree algorithm. The entropy formula is adjusted in basic decision tree algorithm to accommodate multiple labels.

$$-\sum_{j=1}^{q} p(\lambda j)\log p(\lambda j) + q(\lambda j)\log q(\lambda j)$$

Where,

$p(\lambda j) =$ Relative frequency of class λj

$q(\lambda j) = 1 - q(\lambda j)$

Entropy (D) =

B. Multi- Layer neural network (MLNN)

The multi-label neural network's fundamental algorithm is the multi-layer feed forward neural network. Three critical phases are involved in adapting a neural network algorithm for multi-label instance classification:

1. Developing a brand-new error function that encapsulates the features of ML.
2. Adjust the network so that the new error function is minimized.
3. Determining whether an output is in the appropriate set of labels using a threshold function.

C. Back Propagation Multi – Label Learning (BPMLL)

A new global error function that encapsulates the peculiarities of multi-label learning is introduced by BPMLL, which extends the original back-propagation method.

D. Multi- Label K nearest Neighbors (MLKNN)

It is the extension of the K-Nearest Neighbor algorithm. For every label, it applies the KNN algorithm separately. It looks for the cases that are most similar to the test case and takes into account those that are marked as positive and negative. In addition, MLKNN can rank the labels as an output.

E. Multi-Label Boosting (ADABOOST. MH, ADABOOST. MR)

For data with multiple labels, these two techniques are expansions of the fundamental Ada Boost algorithm. Using Ada Boost, hamming loss is decreased. Thanks to AdaBoost.MR, hand accuracy is up.

Table 12: Comparative study of Algorithm Adaptation Methods

Method	Merits	Demerits
C4.5	For dividing decision trees, easier-to-learn and more informative features are utilized.	Does not consider label correlation ship.
MLKNN	Performs better than other algorithms when applied to	It is not possible to classify unlabeled

	picture and text data.	data.
BPMLL	Allows the learning system to be more generalizable.	The training phase is characterized by a high level of neural network complexity.
AdaBoost. MH AdaBoost .MR	Reduced hamming loss and increased accuracy.	It is not possible to classify unlabeled data.

Table 13: Comparative study of MLC methods

Problem Transformation	Algorithm Adaptation
Algorithm Independent	Algorithm Dependent
Multiple model or single Model is used	Single model is used
Data Preprocessing is required	Limited preprocessing Is required.

- **Evaluation Measure**

There is a difference between the evaluation of single-label problems and multi-label problems. Association with multiple labels is common. Experiments make use of four distinct multi-label datasets: Genome, Yeast, Medical, and Scene. The findings shown in the table below demonstrate that, when compared to issue transformation techniques, algorithm adaptation methods have been the best alternative for multi-label approaches.

Classification of an instance may be partially correct or partially incorrect when using more than one label. For multi-label classification issues, the two most common assessment metrics are the example-based measure and the label-based measure.

A. Metrics Based on Examples

Assume that h for (x, Y) predicts a set of labels z=h(x).⊆Consider a multi-label classifier (Leth).

Accuracy:

The percentage of accurate predicted labels relative to the entire number of labels for that instance, including both predicted and real labels, is called accuracy for that instance. The average across all cases is the overall accuracy.

$$\text{Accuracy}(H,D)=\frac{1}{|D|}\sum_{i=1}^{|D|}\frac{|Y_i\cap Z_i|}{|Y_i\cup Z_i|}$$

Precision:

The accuracy rate of the anticipated labels is known as precision.

$$\text{Precision}(H,D)=\frac{1}{|D|}\sum_{i=1}^{|D|}\frac{|Y_i\cap Z_i|}{|Z_i|}$$

The recall measures the proportion of anticipated labels that were accurate.

$$\text{Recall}(H,D)=\frac{1}{|D|}\sum_{i=1}^{|D|}\frac{|Y_i\cap Z_i|}{|Y_i|}$$

Hamming loss:

On average, the number of times that an example's relevance to a class label is wrongly anticipated is reported by Hamming Loss. To calculate hamming loss, we normalize the prediction error (when the wrong label is predicted) and the missing error (when the appropriate label is not predicted) across the entire number of classes and items.

$$\text{Hamming Loss}(H,D)=\frac{1}{|D|}\sum_{i=1}^{|D|}\frac{|Y_i\Delta Z_i|}{|L|}$$

Δ represents the symmetric difference between two sets.

B. Label-Based Assessment

For every label, compute a binary evaluation metric independently. For every label, two averaging processes are utilized: micro average and macro average. The parameters of the confusion matrix are used to calculate the binary evaluation measure.(false positives, false negatives, true positives).

$$\text{M}_{\text{macro}}=\frac{1}{|L|}\sum_{\lambda=1}^{|L|}M(tp_\lambda,fp_\lambda,tn_\lambda,fn_\lambda)$$
$$\text{M}_{\text{micro}}=M(\sum_{\lambda=1}^{|L|}tp_\lambda,\sum_{\lambda=1}^{|L|}fp_\lambda,\sum_{\lambda=1}^{|L|}tn_\lambda,\sum_{\lambda=1}^{|L|}fn_\lambda)$$

7. Results and Discussion

Experiments make use of four distinct multi-label datasets: Genome, Yeast, Medical, and Scene. When comparing algorithm adaptation techniques to problem transformation methods, the findings shown in the table demonstrate that the former is the superior choice for multi-label approaches.

Table 14: Multi label data set statistics

Dataset	Gene base	Yeast	Medical	Scene
#Instances	662	2417	978	2407
#Attributes	1186	103	1449	294
#Labels	27	14	45	6

When comparing the efficacy of various algorithms, it is essential to choose a suitable dataset; multi-label classification issues occur in many real-world fields. To guarantee a thorough assessment across several domains, this work used four unique multi-label datasets. The purpose of compiling gene sequences from different species into the Gene base collection is to make predictions about the functional classes linked to each gene. This dataset presents a significant challenge for multi-label classification algorithms due to its high dimensionality of features (1186 characteristics) and relatively large number of labels.

Predicting the functional classes of genes in the yeast organism is the goal of the Yeast dataset, which is taken from computational biology. Typical of the multi-label classification problems used in bioinformatics research, this dataset has 103 characteristics and 14 labels, which is moderate.

The purpose of the Medical dataset is to link individuals with appropriate illness codes or diagnoses using data gathered from medical case histories. The complexity of medical diagnosis and the possibility of several co-occurring illnesses are reflected in this dataset, which stands out for its high dimensionality (1449 characteristics) and large number of labels.
Last but not least, the Scene dataset is all about picture categorization, with the ability to assign several semantic labels to each image.

Table 15: Experimental result

	Data set	Gene base			Yeast		
Method	Algorithm	H.Loss (%)	Precision (%)	Recall (%)	H.Loss (%)	Precision (%)	Recall (%)
Problem Transformation	BR	0.1	98.9	98.3	24.5	59.9	57.4
	LP	0.2	98.8	97.2	27.9	54.1	53.7
	CLR	0.1	99	98.7	22	65.2	58.4
Algorithm Adaptation	ML-KNN	0.5	99.2	90.1	19.4	72.9	57
	J48ML	0.2	98.9	97.7	28.1	53.3	57.2

Table 15(a): Experimental result

	Data set	Medical			Scene		
Method	Algorithm	H.Loss (%)	Precision (%)	Recall (%)	H.Loss (%)	Precision (%)	Recall (%)
Problem Transformation	BR	1	83.4	78.7	13.7	61.7	62.2
	LP	1.3	77.1	74	14.4	59.8	59.7
	CLR	1	83.6	77.7	13.8	60.6	65.2

Algorithm Adaptation	ML-KNN	1.5	81.2	57.2	8.5	82	67.2
	J48ML	1.3	77.2	74.1	14.4	59.7	60.8

Two problem transformation approaches, Binary Relevance (BR) and Label Power set (LP), one algorithm adaptation method, Classifier Chains (CLR), and two additional methods, ML-KNN and J48ML, are shown in the table as performing various multi-label classification tasks. Using three performance measures, these approaches are tested on four distinct multi-label datasets: Genome, Yeast, Medical, and Scene. Precision, Hamming Loss, and Recall.

1. **Hamming Loss:** As a statistic, Hamming Loss quantifies the percentage of misclassified instance-label pairings. In terms of performance, a smaller number is preferable.
   - Of all the approaches tested on the Genebase dataset, CLR and BR performed the best with the lowest Hamming Loss (0.1).
   - The best performance on the Yeast dataset was seen with CLR (22.0) and ML-KNN (19.4), which had the lowest Hamming Loss.
   - The fact that BR and CLR had the lowest Hamming Loss (1.0) on the Medical dataset suggests that they performed well on this dataset.
   - With the lowest Hamming Loss (8.5) of the algorithms tested on the Scene dataset, ML-KNN is clearly the best option.
2. **Precision:** The percentage of accurately predicted labels relative to the total number of predicted labels is known as precision. It would be ideal to have more precise values.
   - High accuracy values are consistently achieved by BR and CLR across all datasets, suggesting that they reliably anticipate important labels.
   - ML-KNN's 82.0% accuracy rate on the Scene dataset is far higher than that of competing techniques," indicating that it is quite good at predicting labels that are relevant to this dataset.
3. **Recall:** A high recall indicates that a large portion of the genuine labels were properly anticipated. We like recall values that are higher.
   - The maximum recall for the Genebase dataset is 98.3% for BR and 98.7% for CLR, which means that these models can capture the majority of the real labels.
   - It seems that ML-KNN (72.9%) and CLR (65.2%) are the most successful in capturing genuine labels for the Yeast dataset, since they have the greatest recall.
   - The best performance in predicting real labels is shown by BR (78.7%) and CLR (77.7%), which have the greatest recall for the Medical dataset.
   - On the Scene dataset, ML-kNN (67.2% recall) and CLR (65.2% recall) are top performers, indicating that they can accurately capture labels for this dataset.

The attributes, labels, and number of instances in the dataset may affect the performance of multi-label classification algorithms. [27, 28] Techniques for transforming problems, such as BR and CLR, tend to do well on most datasets, whereas approaches for adapting algorithms, such as ML-KNN and J48ML, demonstrate competitive performance on certain datasets. [29, 30] When choosing the best multi-label classification technique, it's crucial to think about the application's needs and the pros and downsides of various performance measures.

8. Conclusion

For multi-label classification, this work included a study of several problem-transformation and algorithm-adaptation techniques. The algorithm adaption approach is the best alternative for multi-label classification compared to the issue transformation method, according to a comparative research and experimental analysis on four datasets: Database, Yeast, Medical, and Scene.

8.1 Findings of the study

This study evaluated various problem transformation methods and algorithm adaptation methods for multi-label classification across four diverse datasets: In the Human Genome, we introduced four terms: the Genebase, yeast as the model organism, the genome of the human applying as a guide in medical practice, and human life is depicted in a chapter. The tests showed that the algorithm adaptation methods, especially ML-KNN, were better than problem transformation methods in some cases. Al-KNN showed up as the best when looking at the pat having errors and exact precision on the scene dataset to be able to predict as well as capture only the relevant labels, therefore applying this method functionalizes well. As well as that, it demonstrated enemy swapping technique, labelling certain categories perfectly like Fungus and Scene. However, problem transformation algorithms like Binary Relevance

(BR) and Classifier Chains (CLR) were the best and most stable algorithms among the different datasets.

8.2 Scope for further research

The area of sequence pattern mining in association rule learning has shown great improvement but there are still some places that need to be explored by using the data mining techniques. The development of techniques and algorithms to solve the problems of high-dimensional, noisy, and uncertain sequential data which can handle (performance) is important. The study of the methods of finding out the complicated and the overlapped patterns could make the use of this technique more universal. In addition, the mixture of (if you'll) domain knowledge and semantic data processing into the mining process balances the interpretability and significance of the patterns. Investigating the parallels and distributed schemes for the pattern mining on big datasets and learning of models is also a promising way.

References

[1] Jiawei Han, Micheline Kamber and Jian Pei, "Data Mining: Concepts and Techniques", Morgan Kaufman publishers is an imprint of Elsevier, 2001

[2] Rakesh Agrawal, and Ramakrishnan Srikant, "Mining sequential patterns". Proceedings of the Eleventh International Conference on Data Engineering, 1995.

[3] M. Chaudhari and C. Mehta, "A Survey on Algorithms for Sequential Pattern Mining", International Journal of Engineering Development and Research, Volume 3, Issue 4, 2015.

[4] Qiankun Zhao and Sourav Bhowmick, "Sequential Pattern Mining: A Survey", Technical Report, CAIS, Nanyang Technological University, Singapore, No. 2003118, 2003.

[5] M. J. Zaki, "SPADE: An Efficient Algorithm for Mining Frequent Sequences", Kluwer Academic Publisher. Machine Learning, 2001, volume 42, pp. 31 -60.

[6] M. Garofalakis, R. Rastogi, and K. Shim, "SPIRIT: Sequential pattern mining with regular expression constraints", 25th VLDB Conference, Edinburgh, Scotland, 1999.

[7] Jay Ayres, Johannes Gehrke, Tomi Yiu, and Jason Flannick, "SPAM: Sequential PAttern Mining using A Bitmap Representation" SIGKDD '02 Edmonton, Alberta, Canada, 2002, ACM 1-58113-567-X/02/0007.

[8] Vishal S. Motegaonkar and Madhav V. Vaidya, "A Survey on Sequential Pattern Mining Algorithms", International Journal of Computer Science and Information Technologies, Vol. 5 (2) , 2014, 2486-2492

[9] J. Pei, J. Han, B. Mortazavi-Asl, H. Pinto, Q. Chen, U. Dayal, and M. C. Hsu, "PrefixSpan: Mining Sequential Patterns Efficiently by Prefix-Projected Pattern-Growth" Proceding 2001 of Internationnal Conference on Data Engineering(ICDE'01), pp. 215-224, Heildelberg, Germany, April 2001.

[10] J. Wang, and J. Han, "BIDE: Efficient mining of frequent closed sequences" Proceedings of the 20th International Conference on Data Engineering, 2004.

[11] Zhenglu Yang and Masaru Kitsuregawa, "LAPIN-SPAM: An improved algorithm for mining sequential pattern" Proceedings of the 21st International Conference on Data Engineering Workshops. 2005, pp. 12-22.

[12] Minh-Thai Tran, Bac Le, Bay Vo, "Combination of dynamic bit vectors and transaction information for mining frequent closed sequences efficiently" Engineering Applications of Artificial Intelligence, Volume 38, February 2015, pp. 183–189

[13] V. PurushothamaRaju, and G. P. Saradhi Varma, "MINING CLOSED SEQUENTIAL PATTERNS IN LARGE SEQUENCE DATABASES" International Journal of Database Management Systems (IJDMS) Volume 7, No.1, February 2015.

[14] Philippe Fournier-Viger, Cheng-Wei Wu, Antonio Gomariz,and Vincent S. Tseng, "VMSP : Efficient Vertical Mining of Maximal Sequential Patterns" Advances in Artificial Intelligence, Volume 8436 of the series Lecture Notes in Computer Science, May 2014, pp. 83-94.

[15] K.M.V. Madan Kumar, P.V.S. Srinivas, and C. Raghavendra Rao, "Sequential Pattern Mining With Multiple Minimum Supports by MS-SPADE" IJCSI International Journal of Computer Science Issues, Vol. 9, Issue 5, No 1, September 2012.

[16] Qian Wang, Darry N Davis, Jiadong Ren, "Mining frequent biological sequences based on bitmap without candidate sequence generation" Computers in Biology and Medicine Volume 69, 2016, pp. 1 52-157

[17] Ling Chen, Wei Liu, "Frequent patterns mining in multiple biological sequences", Computers in Biology and Medicine, Volume 43, 2013, pp. 1444-1452

[18] M K Sohrabi and V Ghods, "CUSE: A Novel Cube-based Approach for Sequential Pattern Mining", 4th International Symposium on Computational and Business Intelligence 2016

[19] P Fournier-Viger, J Chun-W Lin, R Uday Kiran, Yun

Sing Koh, Rincy Thomas, “A Survey of Sequential Pattern Mining”, Data Science and Pattern Recongnition, Ubiquitous Internatitional, Volume 1, Number 1, February 2017.

[20] R Boghey and S Singh, “Sequential Pattern Mining: A Survey on Approaches”, International Conference on Communication Systems and Network Technologies 2013.

[21] J.Wang, J.Han. BIDE: Efficient mining of frequent closed sequences. In Proc. of 2004 Int. Conf. on Data Eng. Apr. 2004, Boston, MA. 79–90.

[22] K.Wang, J.Tan .Incremental discovery of sequential patterns. In Proc of Workshop on Research Issues on Data Mining and Know Discovery. June 1996, Montreal, Canada. 95–102 104.

[23] Lin MY, Lee SY (1998) Incremental update on sequential patterns in large databases. In Proc of the 10th IEEE Int Conf on Tools with Artificial Intelligence. Nov. 1998, Taipei, Taiwan. 24–31.

[24] S.Parthasarathy, MJ.Zaki, M.Ogihara, S.Dwarkadas. Incremental and interactive sequence mining. In Proc of the 8th Int Conf on Information and Know Management. Nov. 1999, Kansas, Missouri, USA. 251–258.

[25] M. J. Zaki, “SPADE: An Efficient Algorithm for Mining Frequent Sequences”, Kluwer Academic Publisher. Machine Learning, 2023, volume 42, pp. 31 -60.

[26] Minh-Thai Tran, Bac Le, Bay Vo, “Combination of dynamic bit vectors and transaction information for mining frequent closed sequences efficiently” Engineering Applications of Artificial Intelligence, Volume 38, February 2022, pp. 183–189

[27] F.Masseglia, P.Poncelet, M.Teisseire. Incremental mining of sequential patterns in large databases., Data and Knowledge Engineering, Vol 46, pp. 97-121, 2003. [

[28] M.Zhang M, Kao B, Cheung D, Yip CL .Efficient algorithms for incremental update of frequent sequences. In Proc of the 6th Pacific-Asia Conf on Know Discovery and Data Mining. May, 2002, Taipei, Taiwan. 186–197.

[29] M.Y.Lin, SY .Lee. Incremental update on sequential patterns in large databases by implicit merging and efficient counting. Information Systems 29: 385-404, 2004.

[30] H.Cheng, X.Yan, J.Han. IncSpan: incremental mining of sequential patterns in large database In: Proceeding of the 2004 ACM SIGKDD international conference on knowledge discovery in databases (KDD’04), Seattle,WA, pp 527–532, 2004.