IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

# Effective Multi-label Classification in Big Data Streams based on K-NN

[1]Htwe Pa Pa Win, [2]Phyo Thu Thu Khine

[1]Associate Professor, [2] Associate Professor

[1, 2] University of Computer Studies, Hpa-an, Myanmar

*Abstract:* Multi-label learning and classification difficulties appeared in many real-world applications of Data mining and AI systems. These situations create a wide range of problems in today's big data era as data streams problems. Although a significant amount of classification for multiple labeling for both normal data and streams data have been made as to the advancement of the basic classifier and different ensemble ways, the multi-label data streams challenge still a hot topic for research works. Therefore, this paper proposes a mechanism to support the big data streams problems of multiple labeling based on lazy classification of K-nearest neighbor, namely ML-KNN. The experiments are carried out using the standard multi-label data streams set and executed with the state-of-the-art data streams algorithms and also compared with the previous multi-label research works. The proposed method makes a significant performance for multi-label classification among the well-known and existing works of different advancements.

***Index Terms* - AI Systems, Big Data, Data Mining, Data Streams, Lazy Classification, ML-KNN, Multi-label Classification**

## I. Introduction

The most real-world applications of today big data era need to classify many different categories in accurate manner, and this emerge multi-label classification problems in data mining paradigms. These problems still exist and newly attracted for research sights from practical world application of image/video annotation, text classification, and social network analysis and many more in various domains. These structured data streams are produced in real time inexpertly with different rates and this create storage problem, processing time and probabilistic distributions of data over time and may be different types of group. Therefore, the application based on data streams classification need to emphasize not only to obtain better performance but also to get the correctly classification among the multi labels [1, 2].

The normal single-label classification is considered that one instance belongs to one class whereas in the multi-label problem, one instance may be allocated to multiple class simultaneously. Since single label classification problem is merely a special one, multiple labelling is considered as a more difficult and complicated case. Furthermore, the multi-label classification for extremely large unlimited data streams which arrived unexpected time is the most difficult problems in data mining fields. This condition become more critical when the data streams of multiple classes has imbalance label problems and the researchers need to give more attention to solve these matters [1, 3].

The multi-label classification problems can be grouped into two main categories algorithm adaptation and problem transformation. In the problem transformation paradigm, multi-label problem is transformed into single label problems and uses single label classifiers and avoids the restrictions of classification mechanism. In the adaptation mechanism, the specified algorithm is modified to handle multi label classification and this may be the most suitable for specific domains but may not be flexible as it may has high complexity [4].

Therefore, this paper proposed a mechanism to solve the multiple class label problems of data streams by using the multi-label K-nearest neighbor, ML-KNN in adaptation manner from basic KNN for the specific domain of text categorization.

The rest of the paper is organized as follow; firstly, section 2 discussed the previous research works to solve the data streams classification problems of multiple labelling. Then section 3 describes the multi-label KNN and section 4 presents about the experimental evolution. Finally, section 5 conclude the multi-label research works.

## II. Related Works

Although, there are many works for the multi label classification problems, this section describes some of the most related researches.

The authors in [2, 5] proposed a model for multi-label classification in adaptively by using AMRule for multi-target regression analysis. They described different streaming methods, issues and various challenges affect the performance of the classification model. Their propose model use aggregator concept by controlling classification of multiple labels with the help of heuristics values. Although, this model provides a high accuracy rate, the time taken is very large and they recommend reducing the time at the pre-processing stage.

The group in [6] propose a modelling method for multi-label data streams classification using recurrent concept. Their proposed framework maintains multi-label concept pool from the changes of incoming instances and corresponding classifiers. They used space dimension reduction method to transform the labels into encoded space and the models is trained in that reduced space. The decoding matrix is updated by the analytical method and it is used to convert the labels to the original space during the testing stage. They used synthetic and real-world datasets to show the effectiveness of their system show the high-performance results.

The wide knowledge of multi-label classification such as techniques, metrics and problem analysis can be found in [7]. The extensive research surveys for multi-label classification for data streams have been done in [8, 9]. They describe the state-of-the-art multi-label algorithms and standard datasets to generate benchmark records for data stream classification.

The researchers in [1] create a novel classification method based on KNN and random walk for multi class data streams, named MLRWKNN. Their method builds random walk graph from vertices set for KNN training samples and the correlations among the training samples are performed as the edge set to reduce the time and space. This method increases the measurement of the similarity using the discrete and continuous features. The label prediction is performed by reducing the subjectivity threshold. They use Flags and Genbase datasets to show the effectiveness of their model and compared among the well-known classifiers and they get the high performance.

The research in [9-12] performs the multi-label data stream classification with different popular well-known algorithms and their various advancement methods and showed their results for the same dataset IMDB used in this paper. Some methods include ensemble mechanisms of different combination of classifiers to improve the performances. When they get the higher accuracy in classification, they suffer from the complexity for time and memory space and vice versa. The detail results will be shown in experiments section. Therefore, the works in this paper intended to try to find the better approach to improve the accuracy result for multi-label streams.

## III. Multi-label K-Nearest Neighbor Classification

K-nearest neighbor, KNN is the simplest approach in classification mechanisms. When the new label arrived, KNN find for nearest neighbor of it by finding the feature distance between the new label and existing label. When the nearest sample is found, it is grouped to the same samples. The KNN only work when new sample comes and any new model is created from KNN, it is so called lazy learning or instance-based Learning [7].

ML-KNN is the advancement of KNN to solve the multi-label problems and developed by [13]. It is not lazy as the classic KNN and builds two pieces of limited information models for each label in the multi-label dataset, MLD; priori probabilities and conditional probabilities.

Let x be an instance and Y be the associated label set of it and $Y \subseteq \mathcal{Y}$, suppose ML-KNN consider KNNs in its implementation. Let the category vector of x be $\vec{y}_x$ where the label component of l, $\vec{y}_t(l)$ $(l \in \mathcal{Y})$ is 1 if $(l \in Y)$ and otherwise it is equal to 0. Let N(x) be the set of classifiers KNNs for x instance in the train set. Therefore, a new membership counting vector is calculated from the neighbors label as follow.

$$\vec{C}_x(l) = \sum_{a \in N(x)} \vec{y}_a(l), \qquad l \in Y \qquad (1)$$

where $\vec{C}_x(l)$ is the number of x neighbors to the class label, $l$.

Let $t$ be the test instance and $N(t)$ be the training set of t. Let $H_1^l$ be the event of instance $t$ with $l$ label while $H_1^l$ be with no label. The event is $E_j^l(j \in \{0,1,\dots,K\})$ among the KNN classifiers of t instance where j instances with label l. Therefore, the category vector $\vec{y}_t$ is examined using the MAP principle as follow.

$$\vec{y}_t(l) = \arg \max_{b\in\{0,1\}} P(H_b^l | E_{\vec{C}_t(l)}^l), \qquad l \in Y \qquad (2)$$

Equation (2) can be written by using the Bayesian rule as follow.

$$\vec{y}_t(l) = \arg \max_{\mathrm{b}\in\{0,1\}} \frac{P(H_b^l)\, P(E_{\vec{C}_t(l)}^l | H_b^l)}{P(E_{\vec{C}_t(l)}^l)}$$

$$= arg \max_{b\in\{0,1\}} P(H_b^l) P(\, E_{\vec{C}_t(l)}^l | H_b^l) \qquad (3)$$

To determine $\vec{y}_t$, the category vector, uses two information model of prior probabilities and posterior probabilities. The prior probabilities is $P\left(H_b^l\right)(l \in y,\ b \in \{0,1\})$ and the posterior probabilities is $P\left(E_j^l \in H_b^l\right)(\, j \in \{0,1,\dots.,K\})$.

## IV. EXPERIMENTAL SETUP AND RESULTS

The comprehensive experiments are carried out to handle the multi-label problems in big data streams to find the best problem solvers. As, the system is intended to solve the big data streams problems of multi-label classification, the dataset for this experiments need to be large dataset and multi-label type. To show the proposed ML-KNN is better than the other multi-label classifiers, the experiments are carried out with different popular corresponding classifiers and compare with the well-known algorithms and previous proposed systems.

### *A. Database Description*

The one of standard multi label dataset, IMDB [14] is used for the experiments. It is the collection of the text description of the movie database of the internet. It consists of 120919 instances, with 1001 binary attributes and 28 class labels.

### *B. Evaluation Metrics*

The most common measurement for multi label streams classification are described.

*Hamming loss*: It is the loss of the average classification error calculated by the fraction of all incorrect labels to the total m labels. The Hamming loss can be defined as follow:

$$\text{Hamming loss } = \frac{1}{\mathrm{N}}\sum_{\mathrm{i}=1}^{\mathrm{N}}\frac{\mathrm{y}_{\mathrm{i}\oplus}\hat{\mathrm{y}}_{\mathrm{i}}}{\mathrm{m}} \qquad (4)$$

where ⊕ is the symmetric difference between the two true label and predicted label sets. The smaller amount of Hamming loss value indicates the better performance of the classifier.

*Hamming Score:* Hamming score is a measure of per-label accuracy defined as:

$$Hamming\ score = \frac{1}{NL}\sum_{i=0}^{N}\sum_{l=0}^{L} I|\ y_l\ = z_l, \qquad y_l \in Y_i\,, z_l \in Z_i \qquad (5)$$

where N is the total instances, L is the total number of labels, Yi is the correct label set and Zi is the predicted label set. The classifier made the final prediction with the highest Hamming Score.

*Accuracy:* The accuracy of the multi-label uses the index of the Jaccard to measure the similarity between the true label set and predict label sets and can be defined as

$$Accuracy\ = \frac{1}{N}\sum_{i=0}^{N}\frac{|Y_i \cap\ Z_i|}{|Y_i \cup\ Z_i|} \qquad (6)$$

*Subset Accuracy*: It is a very strict metric to evaluate the frequency of correctly predicted label set and can be calculated as follow.

$$Subset\ accuracy\ = \frac{1}{N}\sum_{i=0}^{N} I\,|\,Y_i\ = \ Z_i \qquad (7)$$

### *C. Experimental Results*

The experiment is started to find the best classifier among the well-known algorithms of the multi-label classification for the targeted IMDB dataset prescribed above. The algorithms are executed with the prequential multi label evaluation method and they use the default parameters values in the same environment such as the value of k is 5 and with a window, w=100 instances. The performance results for the popular multi-label classification data streams algorithms and the proposed method are described in Table 1.

Table 1. Performance results of well-known multi-label algorithms and proposed mechanism

| Algorithm | Subset accuracy | Hamming Score | Accuracy (%) | Kappa | Kappa Temp | Time (s) | Memory (bytes) |
|---|---|---|---|---|---|---|---|
| ISOUPTree | 0.040 | 0.345 | 0.70 | 0.03 | 0.04 | 80.33 | 60.91 |
| PerceptronClassification | 0.084 | 0.733 | 0.71 | 0.07 | 0.08 | 16.30 | 0.32 |
| RandomMultiLabel | 0.153 | 0.147 | 0.50 | 0.15 | 0.15 | 11.59 | 0.15 |
| NaiveBayes | 0.078 | 0.076 | 0.64 | 0.08 | 0.08 | 27.38 | 0.59 |
| ML-SAM-KNN | 0.158 | 0.145 | 0.82 | 0.14 | 0.15 | 90 | 64.78 |
| Proposed method | 0.206 | 0.199 | 0.86 | 0.20 | 0.21 | 80.88 | 60.06 |

Table 2. Performance Comparison for the Previous Works and Proposed Method

| Previous Works | Hamming loss | Hamming Score | Subset accuracy |
|---|---|---|---|
| iSOUP-MT [10], 2016 | - | 0.9282 | 0.0187 |
| iSOUP-RT [10], 2016 | - | 0.9284 | 0.0026 |
| iSOUP-RT (EBRT) [10], 2016 | - | 0.9286 | 0.0007 |
| iSOUP-MT (EBMT) [10], 2016 | - | 0.9286 | 0.0031 |
| HTPS [10], 2016 | - | 0.8886 | 0.0435 |
| EAHTPS [10], 2016 | - | 0.9151 | 0.1955 |
| Naïve Bayes [11], 2019 | - | 0.0896 | - |
| AODE [11], 2019 | - | 0.1982 | - |
| Mode Imputation [11], 2019 | - | 0.1994 | - |
| Table Expansion [11], 2019 | - | 0.1906 | - |
| ML-SAM-kNN [12], 2019 | 0.0456 | - | 0.154 |
| MHT [12], 2019 | 0.1059 | - | 0.195 |
| MLOzaBag[9], 2019 | 0.1304 | - | 0.134 |
| MLAW[9], 2019 | 0.0986 | - | 0.129 |
| Proposed method | - | 0.199 | 0.206 |

As can be seen from the Table 1, the accuracy of the multi label KNN outperform among the well-known methods. The advancement of ML-KNN and the newest algorithm of KNN, self-adjustment of the memory consumption ML-SAM-KNN give worse results than the normal ML-KNN for the IMDB dataset because the nature of the streams is more suitable for non-drifting. The ML-SAM-KNN consume the longest time and the largest amount of memory for the target dataset. The random multi-label classifier produces the worst result among the tested algorithms. However, it takes only the minimum amount of and memory consumption and time. The ISOUPTree provides the least subset accuracy rate. Then the experiments are continued to compare the previous existing research works and the proposed mechanism. The results for the comparison with the standard measurement are described in Table 2.

The previous works for the IMDB data streams include wide ranges of classification from single classifier to different ensemble classification and many advancement algorithms for the basic classifiers. The results from the Table 2 prove the effectiveness of the proposed method over the previous works and gives the fact that the normal clustering method of KNN for multi label classification, MLKNN is the most suitable for the IMDB streams. It outperforms the most popular fast method of data streams classification, Hoeffding Tree associated classification methods.

## V. CONCLUSION

This paper analysed the various techniques of multi label classification paradigm in data stream classification and proposed a multi label classification method of KNN for classification of data streams in IMDB dataset. The lazy learning results of the statistical information are compared with different well-known algorithms and other previous research works of different advancements. Although ML-KNN utilizes the larger amount of memory to determine the label set, it gives the best result for the IMDB dataset and clearly show the fact that the performance of data streams classifier for the multi-label categorization depend on the nature of data streams.

## REFERENCES

[1] Z. Wang, S. Wang, B. Wan, and W. Wei, “A novel multi-label classification algorithm based on K-nearest neighbor and random walk”, International Journal of Distributed Sensor Networks, Vol. 16, issue 3, March 2020. DOI: https://doi.org/10.1177/1550147720911892

[2] R. Sousa, and J. Gama, “Multi-label classification from high-speed data streams with adaptive model rules and random rules”, Progress in Artificial Intelligence, Vol. 7, pp.177-187, January 2018. DOI: https://doi.org/10.1007/s13748-018-0142-z

[3] K. K. Wankhade, S. S. Dongre, and K. C. Jondhale, “Data stream classification: a review”, Iran Journal of Computer Science, Springer Nature Switzerland AG 2020, May 2020. DOI: https://doi.org/10.1007/s42044-020-00061-3

[4] P. M. El-Kafrawy, A. M. Sauber, and A. Khalil, “Multi-Label classification for Mining Big Data”, Int'l Conf. on Advances in Big Data Analytics, ABDA'15, pp. 75-80, July 2015.

[5] A. M. Alattas, “Adaptive Model over a Multi-label Streaming Data”, IEEE 21st Saudi Computer Society National Computer Conference (NCC), ISBN: 978-1-5386-4111-8, April 2018.

DOI: 10.1109/NCG.2018.8592966

[6] Z. Ahmadi, and S. Kramer, “Modeling Multi-Label Recurrence in Data Streams”, 2019 IEEE International Conference on Big Knowledge (ICBK), pp. 9-16, ISBN: 978-1-7281-4608-9, November 2019. DOI 10.1109/ICBK.2019.00010

[7] F. Herrera, F. Charte, A. J. Rivera, and M. J. Jesus, Multilabel Classification, Problem Analysis, Metrics and Techniques, First Edition, ISBN: 978-3-319-41111-8, pp. 194, Springer International Publishing, 2016. DOI: 10.1007/978-3-319-41111-8

[8] M-L Zhang, and Z-H Zhou, “A Review on Multi-Label Learning Algorithms”, IEEE Transactions on Knowledge and Data Engineering, Vol. 26, Issue 8, ISSN: 1558-2191, pp. 1819-1837, August 2014. DOI: 10.1109/TKDE.2013.39

[9] X. Zheng, P. Li, Z. Chu, and X. Hu, “A Survey on Multi-label Data Stream Classification”, IEEE Access, Vol. 8, ISSN: 2169-3536, pp. 1249-1275, December 2019. DOI: 10.1109/ACCESS.2019.2962059

[10] A. Osojnik, P. Panov, and S. Džeroski, “Multi-label classification via multi-target regression on data streams”, Machine Learning, Vol. 106, pp. 745-770, December 2016. DOI: 10.1007/s10994-016-5613-5

[11] T. T. Nguyen, et. al., “Multi-label classification via label correlation and first order feature dependence in a data stream”, Pattern Recognition, Vol. 90, pp. 35–51, 2019. DOI: https://doi.org/10.1016/j.patcog.2019.01.007

[12] “Efficient Ensemble Classification for Multi-Label Data Streams with Concept Drift”, Information, Vol. 10, No.158, April 2019. DOI:10.3390/info10050158

[13] M-L. Zhang, and Z.H. Zhou, “ML-KNN: A lazy learning approach to multi-label learning”, Pattern Recognition, Vol. 40, No. 7, pp. 2038-2048, 2017. DOI: https://doi.org/10.1016/j.patcog.2006.12.019

[14] https://sourceforge.net/projects/meka/files/Datasets/