



Cost based Decision Tree for Sequence Data Classification

¹Dr. D. Mabuni, ²Dr. S. Aquter Babu

¹Assistant Professor, ² Professor

Dept. of Computer Science and Technology

Dravidian University, Kuppam, Andhra Pradesh, India

Abstract: Sequence data classification is an important task in machine learning. The decision trees C4.5 and CART are useful for classifying fixed length records but not efficient to classify sequence data directly. Sequence data records are variable length records as a result of this existing classification techniques are not applicable directly. In this context there is a need to investigate for finding better techniques which are convenient for classifying sequence data directly by using C4.5 and CART decision trees. In the present paper, a new technique called cost based decision tree for sequence data classification directly is proposed. A sample sales details database is used as the training dataset. Proposed direct sequence data classification technique is applied on this training dataset. Cost based measure is computed for the data present in each internal node of the decision tree and this measure is used for internal node split. The proposed cost based decision tree metric for sequence data classification is easy to understand, apply and convenient for classifying sequence data directly. Finally, cost based decision tree for sequence data classification is constructed for the sample training dataset.

Index Terms: Sequence data classification, average cost, total cost, vector data, fixed length records, variable length records, additive tree, machine learning, data mining.

I. INTRODUCTION

Data classification is a fundamental task in many machine learning applications. C4.5 is the standard decision tree classifier used for data classification. It is occupied in the top-1 position among the top-10 positions of the data mining algorithms list. After invention of C4.5 many improvements have been made for effective data classification. Its main advantages are interpretability, comprehensibility, benchmarking capability, ability to manage large data with two class labels as well as multi-class labels. Decision tree classification technique is a supervised learning technique because class labels are given for all the tuples in the training dataset.

Data classification techniques are frequently used in many machine learning applications. Various popular data classification techniques are decision trees (C4.5), classification and regression trees (CART), neural networks, support vector machines (SVM) and K-nearest neighbor classifiers. Sometimes the interpretability of the classification data model is more important than the accuracy of the classification algorithm. Applications of C4.5 and CART decision trees are - health care, business, medical diagnoses, finance, retail, recommendation systems, research, business, agriculture and medical data gene sequence analysis. It is a known fact that sequence data classification is an important task in machine learning.

The availability of sequence data is abundant in many real world applications. This necessitates the immediate need of finding efficient, effective, accurate, fast, robust and better sequence data classification techniques. The existing sequence data related literature says that sequence data classification plays a major role in many real time applications in the diversified domains. Sequence data are generated in many applications of science, engineering, technology, and medical domains.

II. RELATED WORK

Machine learning applications development and usage are increasing every day. Data classification is one of the most important applications of machine learning. There exist numerous classification techniques in machine learning. Out of the many classification techniques the C4.5 and CART are predominantly being used in almost all real time applications because those classifiers are interpret-able, comprehensive, benchmark, standard, general, popular, well known and easy to understand and

implement. In the data mining history the C4.5 classification algorithm has occupied top number one position among the top-10 positions.

Research on C4.5 and CART are continuously progressing with positive approach only and these decision tree algorithms are placed in the standard bench marking level and ever green data classification framework. Z. He et al. [1] said that C4.5 and CART (classification and regression tree) algorithms are popular classification and regression algorithms but they work only on fixed length vector data and they don't work on sequence input data. The length of vector is fixed but the length of sequence data is not fixed but it is a variable one in many cases. One simple strategy to process sequence data is first convert it into vector data and then use already existing classifiers for data classification. But converting sequences into discrete features is a tedious task and some times impossible too. Even when sequences are converted into discrete features there is a possibility to miss some important features. This necessitates the need for finding direct new techniques for effective sequence handling and optimal management. Authors proposed a new tree-based algorithm for sequence data classification. Experimental results have shown that better performance results are obtained when compared with advanced and state-of-the-art algorithms.

Z. Xing et al. [2] said that classification of discrete sequences is a more challenging task than the classification of vector data. To handle the sequence classification problem various methods have been proposed. For very large datasets, creating a classifier with good generalization ability by converting sequences into vectors is a tough task. Some times ordering of elements in the sequences is important. Incorporating sequential information details into the predictive model information is really a difficult task. Existing sequence data classification methods are divided into three types- feature based, distance based and model based methods. Y. Hamuro et al. [3] said that classification of sequence data is one of the important problems now a days and they have developed a decision tree package for classifying sequence data by using variables. Proposed method includes procedures for handling sequence data and also it contains a special index for classification purpose.

A. Brunello et al. [4] proposed a new algorithm called J48S, which is like a C4.5 decision tree, for sequence data classification. Authors said that sequence data management is necessary in many business applications. H. Su [5] proposed sequential pattern mining algorithm called representative sequential patter mining, which is suitable for multi-label classification methods. This new decision tree is called multi-valued and sequence-labeled decision tree. N. Lesh [6] proposed a new sequence data classification algorithm, which initially transforms sequences into fixed length vectors and then existing classifiers can be used for sequence classification. T. P. Exarchos et al. [7] proposed a sequence data classification algorithm that executes in two steps. In the first step sequential data model is defined by assigning weights. In the second step optimization method is applied for finding optimal weights for achieving optimal classification accuracy.

C. Zhou et al. [8] proposed pattern based sequence classification method where patterns are represented in the form of rules. Desired patterns are measured by using cohesion and support metrics. First patterns are identified and then identified patterns are transformed into rules. Presented two ways of creating sequence data classifiers. E. Eggho et al. [9] said that pattern based sequence data classification is really a complex task and in sequence data classification they have identified two important problems in classifying sequence data patterns. These problems are parameter tuning and instability of desired patterns. Authors have proposed parameter free methods for sequence classification. D. Nguyen et al. [10] identified sparsity and high dimensional problems in pattern data mining. To overcome these problems authors proposed an unsupervised method for sequence data classification.

Ifrim and Wiuf [11] proposed a framework for discriminating sequence data classification. The performance of the proposed method is comparable to many of the advanced methods for sequence data classification. Nguyen et al. [12] clearly explained about many of the existing sequential data classifiers. Various smoothing techniques are applied for sequential data classifiers and then the results are noted down. Finally they have identified that smoothing techniques have improved sequence data classification accuracy. Z. Zheng et al. [13] said that only little work has been done in contrast sequential pattern(CSP) mining. As a result of this authors proposed contrast sequential pattern mining approach by using a special and effective data structure called CSP-tree. Also proposed many ideas to reduce the search space. Experimental results have shown that performance of the proposed method is superior on many datasets.

Z. He et al. [14] said that discriminating sequential pattern mining is crucial in many applications but they contain many false positives. To overcome this problem authors have proposed significant-based discriminating sequential pattern mining. Experimentally verified that performance of the proposed technique is very high. Z. He et al. [15] said that discriminating sequential pattern mining applications are abundant. Recently many such discriminating pattern mining techniques are proposed but their main disadvantage is that there exists many redundant patterns. To alleviate this problem authors have proposed a new algorithm called conditional discriminating pattern mining. Authors [16] said that sequence groups comparison plays a major role in many applications and they proposed emerging sequences as sub-sequences and these sub-sequences are frequent in group and in frequent in the other group. To overcome many of the existing problems authors proposed a suffix-tree based framework.

J. De. Smedt et al. [17] proposed a new algorithm for sequence data classification and this algorithm is called behavioral constraint miner algorithm. Authors [18] proposed a new algorithm for mining frequent closed sequences without candidate maintenance. Sequential data mining is an inevitable data mining task with innumerable applications such as medical analysis, market analysis, protein analysis, customer analysis, transaction analysis etc. J. M. Luna et al. [19] proposed a new classification algorithm called additive tree which is equivalent to CART in many respects and uses only a single parameter. Additive tree is designed and developed for high interpretability and high predictive performance.

Rokach et al. [20] examined thoroughly the effects of the operation sequence on the quality of the product by using different types of machine learning techniques. A novel framework has been developed for extracting sequence patterns of any length. The proposed algorithm consists of 4 steps and in the last step a decision tree is used for sequence classification. N. R. Mabroukeh and C. I. Ezeife [21] said that different types sequential pattern mining algorithms have been proposed in the literature. These existing sequential pattern mining algorithms have been modified for representation and management of closed, maximal, incremental and hierarchical sequences. Authors [22] said that sequential data mining applications usage is rapidly increasing day by day. Distributed sequential data mining is also increasing in use. Parallel sequential data mining algorithms are also using in many real time applications. Authors have reviewed many of these algorithms for making many possible improvements.

III. PROBLEM DEFINITION

Standard classification algorithms are available for classifying fixed length records. These fixed length records are called vectors. Almost all existing classification techniques are able to classify vector data but not able to classify variable length records such as sequences. Sequence data are predominantly variable in length by default. Main problem is that separate advanced sequence data handling techniques are need for sequence data classification and the techniques must be efficient, effective, robust and easy to use and implement.

IV. Proposed Method for Cost based Decision Tree for Sequence Data Classification

In the present paper a new technique which is cost based technique is proposed for sequence data classification using decision tree method. A sample set of tuples of items belonging to the sales data details are taken as the training dataset. During the process of creation of proposed decision tree, at each internal node of the decision tree a split value is to be computed and it is used for each internal node data splitting. If all the tuples in the root node belongs to the same class label then a leaf node is created with major class label. If tuples contain both the class labels then the average cost of all the tuples in the internal node is computed and this average cost is greater than the threshold value set by the company then node split is applied and data is partitioned into two parts; otherwise if the average cost is less than the threshold value set by the company then a leaf node is created. Same procedure is applied for each internal node of the decision tree until the completion of the decision tree construction.

Root node is an example for internal node and at the very beginning tuples in the root node are processed if all tuples belongs the same class label a leaf node is created; otherwise at the root node all the sequence data tuples are considered and the average cost of all transaction tuples is computed at the root node. This average cost is taken as split node value and if this value is greater than the threshold value set by the company then tuples are partitioned based on the computed average split value. Tuples whose individual normal score is greater than or equal to the average cost are placed in the left partition and all the tuples whose scoring cost is less than the average cost are placed in the right partition. This process is repeated until all the tuples are classified correctly or average scoring cost of the node is less than the specified or set threshold value. A leaf node is created if all the tuples belongs to the same class label or average scoring cost is less than specified threshold cost.

4.1 Sample Dataset

Database of sales details of items of a particular company is taken into consideration for easy explanation of purchased sequences of items. Suppose the list of distinct sales items available in the ABCD company is shown in the Table-1.

Table-1. Master Database of Distinct Items of ABCD Company

S.No	Item Name	Cost
1	Computer	50,000
2	Tablet	40,000
3	Plotter	20,000
4	Printer	20,000
5	CPU	10000
6	Ram	9000
7	Monitor	8000
8	Hard Disk	6000
9	Tape Drive	5000
10	Pen Drive	4000
11	Mouse	3000
12	Key board	2000
13	Remote	1000
14	Light Pen	2000
15	Floppy	1000
16	CD	1000

Now consider a training dataset of items of sequences of transactions and their details as shown in the Table-2. For the easy understanding purpose, cost in K (thousands) and total cost columns are given in the Table-2. This training dataset contains total of sixteen tuples of which eleven tuples belongs to yes class label and five tuples belongs no class label. How tuple costs are computed is shown in the column cost in K.

Table-2 Sample Training Dataset

Transaction No	Items purchased	Class	Cost in K	Total Cost in K
T1	Computer, Tablet, Plotter, Printer	yes	50 + 40 + 20 + 20	130
T2	Computer, Plotter, Printer	yes	50 + 20 + 20	90
T3	Printer, Plotter	yes	20 + 20	40
T4	Computer, RAM, CPU, Plotter	yes	50 + 9 + 10 + 20	89
T5	Tablet, Printer	yes	40 + 20	60
T6	CPU, RAM, Mouse	no	10 + 9 + 3	22
T7	RAM, CPU, Key board, Printer	yes	9 + 10 + 2 + 20	41
T8	CPU, RAM, Key board, Monitor, Mouse	no	10 + 9 + 2 + 8 + 3	32
T9	Computer, CPU, RAM	yes	50 + 10 + 9	69
T10	Pen drive, CD, Mouse, Tablet	yes	4 + 1 + 3 + 40	48
T11	Mouse, pen drive, Key board, Monitor, Computer	yes	3 + 4 + 2 + 8 + 50	67
T12	CPU, RAM	no	10 + 9	19
T13	Monitor, Pen drive	no	8 + 4	12
T14	RAM, Mouse, Key board, CD, Computer	yes	9 + 3 + 2 + 1 + 50	65
T15	RAM, CPU, Mouse	no	9 + 10 + 3	22
T16	Computer, Tablet, Printer, CPU	yes	50 + 40 + 20 + 10	120
Average cost = 52.875 K				

4.2 Average Cost Computations at Each Internal Node of the Decision Tree

In many situations, in many contexts of business conditions, the average value plays an important role in many applications. During sequence decision tree creation process at each internal node average cost of all tuples is computed. If all tuples in the internal node have the same class label or average cost is less than the specified threshold then the node is converted into leaf node; otherwise data in the current internal node is split into left and right partitions and decision tree grows. Same procedure is applied for each internal node of the sequence decision tree. Here, average cost of the internal node is taken as a measure of split. Other types of split measures are also available for internal node split. In this paper, average cost of all tuples in the current internal node is taken on the fly as splitting value.

At the very beginning average cost of the entire training dataset (root node cost) is 52.875 K and average cost of the root node of the sequence decision tree is taken as 52.875 K. Suppose threshold transaction cost set by the company is 40 K. For each internal node average tuple cost is computed on the fly. Initial threshold cost is fixed by the company but the average tuple cost varies from one internal node to other internal node. A simple linear cost metric is used to convert each transaction into its total cost. Other types of tuple conversion metrics are also possible.

4.3 Algorithm: Cost based Decision Tree for Sequence Data Classification

Input: Training Dataset, T

Output: cost based decision tree for sequence data classification

CBDT_sequence(node T)

1. if all the tuples belongs to the same class label then
2. create a leaf node with appropriate class label
3. return
4. endif
5. total_cost = 0
6. for each tuple in the currently selected internal node, T, do
7. compute tuple_cost in K
8. total_cost = total_cost + tuple_cost
9. endfor
10. compute aggregate cost of currently selected internal node.
11. aggregate_cost = total_cost / size of internal node
12. if aggregate_cost >= threshold_cost then
13. for each tuple i in the internal node do
14. if cost(T_i) >= threshold_cost then

```

15. place tuple  $T_i$  in the left partition,  $T_{left}$ 
16. else
17. place tuple  $T_i$  in the right partition,  $T_{right}$ 
18. endif
19. endfor
20. call CBDT_sequence(node  $T_{left}$ )
21. call CBDT_sequence(node  $T_{right}$ )
22. else // if part of step 12 continuation
23. create leaf node
24. endif // if part of step 12

```

4.4 Explanation of the cost based decision tree for sequence data classification

Line numbers 1, 2 and 3 are executed when the selected internal node contains all tuples with only single class label. In this situation a leaf node is created immediately.

Line numbers 5 to 11 are for creating the total cost and then the aggregate cost of all tuples present in the currently selected internal node.

Line numbers 12 to 19 are for creating the data tuples in the current internal node into two partitions- left partition and right partition if the average cost of the current internal node is greater than or equal to the threshold value set by the company.

Line number 20 is for calling the left partition recursively

Line number 21 is for calling the right partition recursively

Line number 23 for creating the leaf node with appropriate class label when average cost of the current internal node is less then the threshold cost of the company.

4.5 Created Sequence Decision Tree for the given training dataset

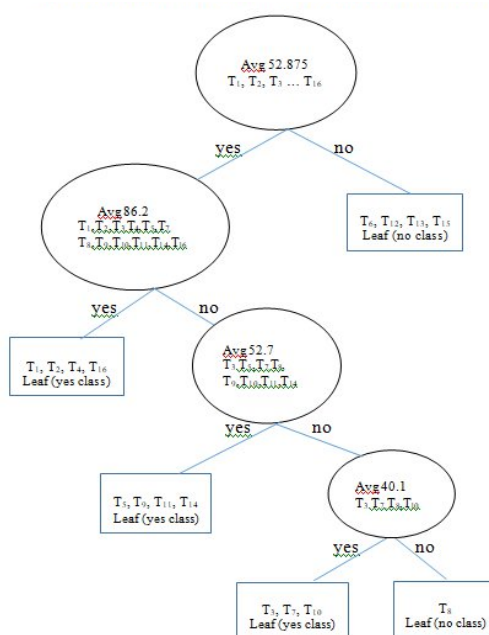


Figure-1 Sequence Decision Tree for the Training dataset show in Table-1

Pruning is a powerful technique available in machine learning. There are two types of pruning techniques. The first one is called pre-pruning and the second one is called post pruning. In the present context ppost pruning is applied for the generated cost based decision tree for sequence data classification. Post pruning is applied after complete sequence decision tree is created. Post pruning proceeds from bottom to top of the sequence decision tree. Post pruned sequence decision tree is shown in Figure-2.

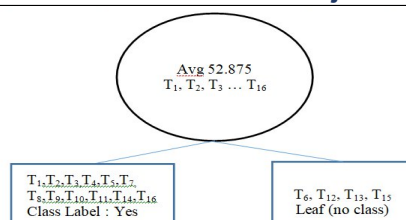


Figure-2 Post Pruned Decision tree for sequence data classification

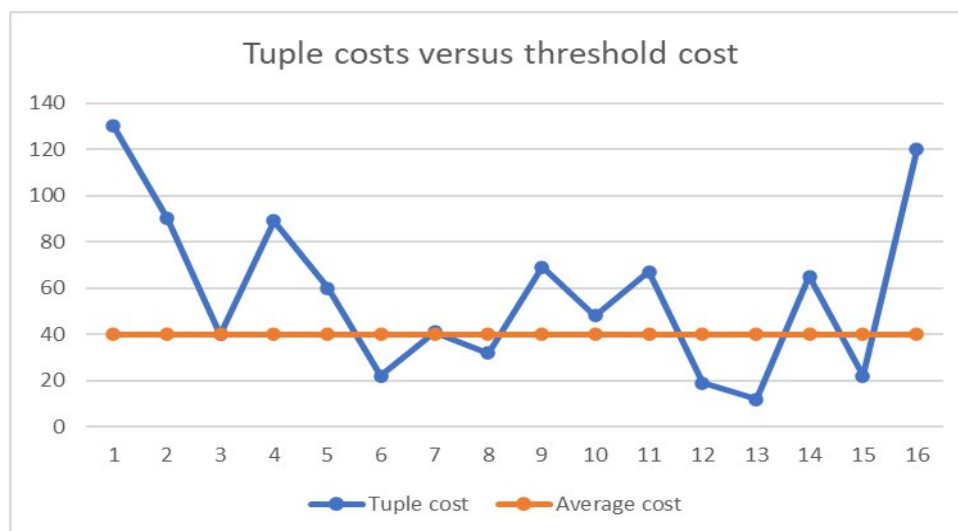


Figure-3 Tuple cost versus threshold cost for the root node

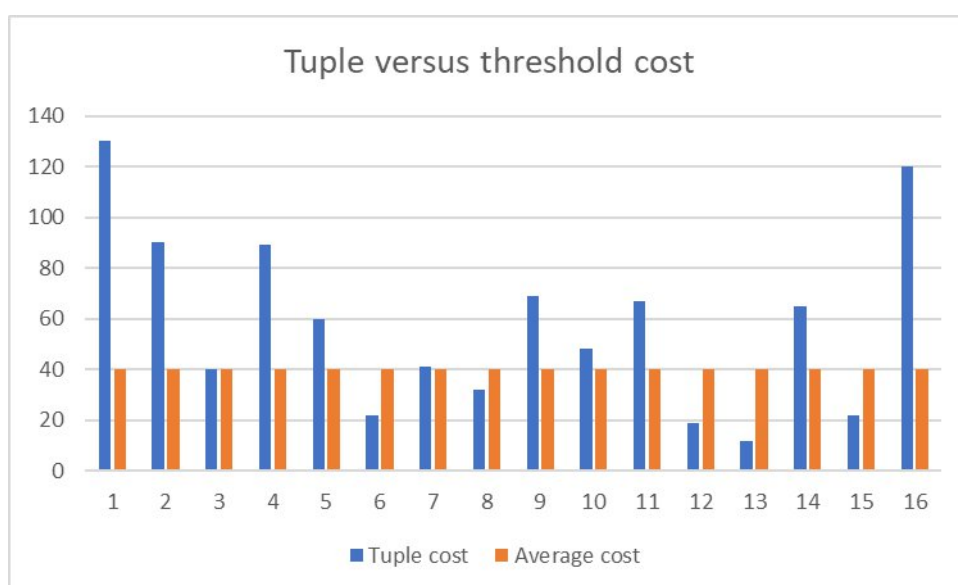


Figure-4 Tuple cost versus threshold cost for the root node

Conclusions

Classification is a primary and very important task in both data mining as well as in many machine learning applications. Classification is a supervised learning techniques where class labels are given for all the training tuples in the training dataset. Existing classification techniques such as C4.5 and CART (classification and regression trees) have the limitation that they can classify only vector data but not sequence data. But sequence data classification applications are rapidly increasing day by day and their demand is increasing exponentially. So, there is an urgent need to invent advanced sequence data classification techniques in machine learning. Decision tree for sequence data classification is one new and advanced technique which is propose recently. In this context in the present paper a new cost based decision tree technique for sequence data classification is proposed and applied on the selected simple sales data details training dataset and the resulted decision tree is explained in detail. In the future, special care and strong efforts will be applied and focused on finding better than the state of the art sequence data classification techniques.

References

- [1] Z. He, Z. Wu, G. Xu, Y. Liu and Q. Zou, "Decision Tree for Sequences", IEEE Transactions on Knowledge and Data Engineering, Volume: 35, Issue: 1
- [2] Z. Xing, J. Pei and E. Keogh, "A brief survey on sequence classification", ACM SIGKDD Explorations Newslett., vol. 12, no. 1, pp. 40-48, 2010.
- [3] Y. Hamuro, M. Nakamoto, S. Cheung and E. H. Ip, "mbonsai: Application package for sequence classification by tree methodology", J. Statistical. Software., vol. 86, no. 6, pp. 1-30, 2018.
- [4] A. Brunello, E. Marzano, A. Montanari and G. Sciacicco, "J48S: A sequence classification approach to text analysis based on decision trees", Proc. 24th Int. Conf. Inf. Softw. Technol., pp. 240-256, 2018.
- [5] C.L. Hsu, "A multi-valued and sequential-labeled decision tree method for recommending sequential patterns in cold-start situations", Appl. Intell., vol. 51, no. 1, pp. 506-526, 2021.
- [6] N. Lesh, M. J. Zaki and M. Ogihara, "Mining features for sequence classification", Proc. 5th ACM SIGKDD Int. Conf. Knowl. Discov. Data Mining, pp. 342-346, 1999.
- [7] T. P. Exarchos, M. G. Tsipouras, C. Papaloukas and D. I. Fotiadis, "A two-stage methodology for sequence classification based on sequential pattern mining and optimization", Data Knowl. Eng., vol. 66, no. 3, pp. 467-487, 2008
- [8] C. Zhou, B. Cule and B. Goethals, "Pattern based sequence classification", IEEE Trans. Knowl. Data Eng., vol. 28, no. 5, pp. 1285-1298, May 2016.
- [9] E. Egho, D. Gay, M. Boullé, N. Voisine and F. Clérot, "A user parameter-free approach for mining robust sequential classification rules", Knowl. Inf. Syst., vol. 52, pp. 53-81, 2017.
- [10] D. Nguyen, W. Luo, T. D. Nguyen, S. Venkatesh and D. Phung, "SqN2Vec: Learning sequence representation via sequential patterns with a gap constraint", Proc. Joint Eur. Conf. Mach. Learn. Knowl. Discov. Databases, pp. 569-584, 2018.
- [11] G. Ifrim and C. Wiuf, "Bounded coordinate-descent for biological sequence classification in high dimensional predictor space", Proc. 17th ACM SIGKDD Int. Conf. Knowl. Discov. Data Mining, pp. 708-716, 2011.
- [12] T. Le Nguyen, S. Gsponer and G. Ifrim, "Time series classification by sequence learning in all-subsequence space", Proc. IEEE 33rd Int. Conf. Data Eng., pp. 947-958, 2017.
- [13] Z. Zheng, W. Wei, C. Liu, W. Cao, L. Cao and M. Bhatia, "An effective contrast sequential pattern mining approach to taxpayer behavior analysis", World Wide Web, vol. 19, no. 4, pp. 633-651, 2015.
- [14] Z. He, S. Zhang and J. Wu, "Significance-based discriminating sequential pattern mining", Expert Syst. Appl., vol. 122, no. 15, pp. 54-64, 2019.
- [15] Z. He, S. Zhang, F. Gu and J. Wu, "Mining conditional discriminating sequential patterns", Inf. Sci., vol. 478, pp. 524-539, 2019.
- [16] K. Deng and O. R. Zaïane, "Contrasting sequence groups by emerging sequences", Proc. 12th Int. Conf. Discov. Sci., pp. 377-384, 2009.
- [17] J. De Smedt, G. Deeva and J. De Weerd, "Mining behavioral sequence constraints for classification", IEEE Trans. Knowl. Data Eng., vol. 32, no. 6, pp. 1130-1142, Jun. 2020.
- [18] J. Wang, J. Han, "BIDE: efficient mining of frequent closed sequences", In Twentieth IEEE International Conference on Data Engineering (ICDE 2004), pp. 79-90 (2004)
- [19] J. M. Luna et al., "Building more accurate decision trees with the additive tree", Proc. Nat. Acad. Sci. USA, vol. 116, no. 40, pp. 19887-19893, 2019.
- [20] L. Rokach, R. Romano and O. Maimon, "Mining Manufacturing Databases to Discover the Effect of Operation Sequence on the Product Quality." 2008, Journal of Intelligent Manufacturing, 19(3), 313-325. doi:10.1007/s10845-008-0084-6.
- [21] N. R. Mabroukeh, C. I. Ezeife, "A taxonomy of sequential pattern mining algorithms". ACM Comput. Surv. 43(1), 1-41 (2010)
- [22] P. Fournier-Viger, J.C.W. Lin, R. U. Kiran, Y.S. Koh, R. Thomas, "A survey of sequential pattern mining". Data Sci. Pattern Recognit. 1(1), 54-77 (2017)