

We are IntechOpen, the world's leading publisher of Open Access books Built by scientists, for scientists

6,900

Open access books available

185,000

International authors and editors

200M

Downloads

Our authors are among the

154

Countries delivered to

TOP 1%

most cited scientists

12.2%

Contributors from top 500 universities



WEB OF SCIENCE™

Selection of our books indexed in the Book Citation Index
in Web of Science™ Core Collection (BKCI)

Interested in publishing with us?
Contact book.department@intechopen.com

Numbers displayed above are based on latest data collected.
For more information visit www.intechopen.com



Association Rule Mining on Big Data Sets

*Oguz Celik, Muruvvet Hasanbasoglu, Mehmet S. Aktas
and Oya Kalipsiz*

Abstract

An accurate, complete, and rapid establishment of customer needs and existence of product recommendations are crucial points in terms of increasing customer satisfaction level in various different sectors such as the banking sector. Due to the significant increase in the number of transactions and customers, analyzing costs regarding time and consumption of memory becomes higher. In order to increase the performance of the product recommendation, we discuss an approach, a sample data creation process, to association rule mining. Thus instead of processing whole population, processing on a sample that represents the population is used to decrease time of analysis and consumption of memory. In this regard, sample composing methods, sample size determination techniques, the tests which measure the similarity between sample and population, and association rules (ARs) derived from the sample were examined. The mutual buying behavior of the customers was found using a well-known association rule mining algorithm. Techniques were compared according to the criteria of complete rule derivation and time consumption.

Keywords: big data, sampling, association rule mining, data mining, data preprocessing techniques

1. Introduction

Thanks to improved storage capacities, databases in various fields such as banking have grown up to a rich level. Most of the strategic sales and marketing decisions are taken by processing these data. For example, strategies such as cross-sell, up-sell, or risk management are being created as a result of processing the customer data. Because of the increasing number of customers and the need for a higher processing capacity, it has made it more difficult to identify the customer requirements in a rapid and accurate way and to present solution recommendations. Innovative data mining applications and techniques are required to solve this issue [1].

The market basket analysis is one of the data mining methods applied to identify the pattern which is found in product ownership data of customers. Thanks to this analysis, a pattern among the products frequently bought together by the customers can be established. The obtained pattern plays an active role in developing cross-sell and up-sell strategies.

Market basket analysis consists of two main processes. These are clustering and association processes, respectively. The clustering process involves grouping of similar customers in terms of clusters. Thus, those customers which should be

examined in the same category will be identified. During the association process, commonness in buying behavior of customers through a selected cluster is being identified, assuming that clustered customers having similar characteristics would demonstrate similar buying behaviors.

As the banking databases have grown up to a very high volume, the association process has become a very costly process in terms of time and memory consumption. In order to improve the time and memory performance, sampling process should be included in the previous phase of association.

In this regard, a sample which involves less observations in comparison to the whole data is used. We use the term “space” to refer the whole data set. In case the representation capability of the obtained sampling is high, loss of data is minimized, and the association process is realized through the sample instead of the space itself. Thus, less data shall be processed, and association rules (ARs) shall be obtained faster by consuming less memory.

As the subject of this book chapter was focused onto the banking data, customer segmentation conducted by the bank data was accepted as the clustering. As a result of the segmentation, clusters created by similar customers were used as input of sampling.

In this chapter, sample creation methods, techniques to find ideal sampling size, the space representing capability of these samples generated by these techniques, and association rules discovered through these samples were examined, respectively. Association rules obtained from both the space and sample were used to verify the sampling process. Besides, the spared amount in terms of time consumption was calculated.

This book chapter was organized as follows: Section 2 explains the studies toward deriving association rules through the space and sampling. Section 3 explains the parameters required to obtain association rules and the Apriori algorithm. Section 4 contains parameters to create the sample, sample creation methods, and the techniques used to calculate the sample size. Section 5 examines association rules obtained from the space and the sample and the results showing the representation capability of the sample for the respective space and the results showing rewards in terms of time consumption. Section 6 gives an overview and concludes the chapter.

2. Related work

In association rule mining, first the item sets, which are found together frequently, are found, and then the rules are obtained from these item sets.

Association algorithms are classified according to characteristics of the obtained item sets. In early studies Agrawal-Imielinski-Swami (AIS) algorithm which was allowed to find wide item sets was used, and then algorithms were found such as Apriori, which were used frequently now and which were able to process the bigger data sets faster [2].

The mutual usage of association discovery and sample creation methods is not a new approach. Sample creation studies toward association detection have begun with papers demonstrating mathematically that it was possible to create a sample which maintained the characteristics of the space. The following studies involved several techniques calculating the optimal number of observations [3–7].

At the beginning of the sample size detection studies, the data to be sampled were not considered; they have tried to determine the sample size using parameters not depending on the data such as margin of error, minimum support, and minimum confidence [3]. In current studies, formulas (using variables such as maximal process length or Vapnik-Chervonenkis (VC) size of the data cluster) considering the data characteristics have appeared [4–7]. There exists a number of studies focusing on how the management of metadata of big data sets are provided in a distributed

computing setting [8–11]. Moreover, there exists a number of studies that are conducted in the field of information systems for managing distributed data storage platforms [12–16]. Unlike these studies, this chapter focuses on extracting meaningful information, i.e. association rules, from the big data sets. Initial results of the experimental studies, covered in this chapter, were reported in a previous study [17]. Sections 3 and 4 give detailed information on association detection and sample creation methods, respectively, and explain the techniques used in this study.

3. Association detection methods

In data mining, it is used to determine the pattern found among the association algorithms and observations [2, 18, 19]. In case any organization's transaction database is discussed, an analogy can be established between the observations and customers and between areas where a pattern is tried to be found and the bought products. Patterns obtained by association algorithms are processed to obtain association rules.

Association rules may be defined as follows: let us call each subset of products within the database an "itemset," and let us call each set of products purchased together by the customer a "transaction." The support count of any itemset is defined as the number of transactions associated with the items in the set within the database. The support indicates the ratio of support count to the number of transactions within the database. The itemset which meets the minimum support requirement is called the frequent itemset (FI).

For example, if a database with 10 transactions contains product A in 3 different transactions, then the product A's support count is 3, and its support is 0.3. In case the minimum support is defined by a value lower than 0.3, then the product A will be classified as FI.

There are several algorithms deriving FI using the transactions within the database [2, 18]. In this chapter, Apriori algorithm was preferred due its ability of deriving all itemsets within the space. This algorithm derives primarily candidate itemsets starting with one-element itemset from the database. Those providing minimum support from candidate itemsets are filtered and recorded as FI. New candidate itemsets are created from the FI obtained in the previous step by increasing the number of elements. In each step, the candidate itemsets are passed through a minimum support test, and the algorithm continues until no FI with k-elements can be generated.

Among the elements of the FI obtained from the database, it is possible to derive association rules in $A \rightarrow B$ format. Then, AR's support gets equal to AUB itemset support. The confidence is defined as the ratio of AUB itemset support to the A itemset support. AR should meet the minimum confidence requirement specified by the customer [2].

Assuming that $A \rightarrow B$ rule has a support of s and the confidence of c , we can derive that the itemsets A and B in the whole database are associated with a probability of s and a customer owning the itemset A might be an owner of the itemset B with a probability of c .

To find out all ARs within the database, a rule mining algorithm is applied to each FI obtained. Candidate rule combinations are created for rule mining among all subsets of a selected FI in $A \rightarrow B$ format. Those providing minimal confidence from candidate rules are filtered and recorded as association rule.

4. Sample creation methodology

Sample creation is the process of creating a subset containing the characteristics of a data set. The subset created through sampling is expected to represent the

data set (space). In traditional statistical methods, the similarity of two data sets is measured by either χ^2 test or Kolmogorov–Smirnov (K-S) test.

In this study, these tests were utilized, in order to measure the similarity of the created sample in comparison with its space. A comparison was conducted through p values (the probability p of finding the space characteristics) of the statistics resulting from both tests. In case the obtained p value exceeds 0.05, it can be deducted that “the sample is similar to the space with a probability of at least 95%.”

Sample creation is discussed under two topics, i.e., sample creation methods and sample size determination techniques. Sample creation methods are explained in Chapter 4.1 and sample size determination techniques explained in Chapter 4.2.

4.1 Sample creation methods

When creating samples from the space, it is possible to use several sample creation methods. These methods are classified according to the selection of observations from the space. The main sample creation methods are as follows:

4.1.1 Simple random sampling

The observations within the space are selected without following a specific routine. The selection probability of each observation is equal.

Systematic sampling: The observations within the space are numbered. Sampling interval is created by dividing the space size to the observation size. A random number is selected. The observation sample at this number from each interval is included.

4.1.2 Stratified sampling

This is used where the observations within the space can be divided into groups. The samples are created maintaining the ratio between the number of observations of groups within the space and the total number of observations. The selection probability of each observation in the same stratus is equal.

4.1.3 Cluster sampling

This is used where the observations within the space can be divided into groups. After the groups are determined, they are selected using the simple random sampling method. All observations within selected groups are included into the sample.

4.1.4 Multistage sampling

This is used where the observations within the space can be divided into groups. After the groups are determined, groups are selected by the simple random sampling method. Unlike cluster sampling, observations to be selected from groups are determined by the simple random sampling method.

Among the mentioned methods, the simple random sampling method stands up by its high speed. As the methods, which require creation of groups within the space and sorting of observations, need a pre-analysis, their time consumption is more than the simple random sampling method.

4.2 Sample size determination methods

The expected parameter in sample creation methods is the size of the sample to be created. When the optimal sample size is calculated, a number which will not decrease

its space representing capability should be found. Under association detection algorithms, it is important to derive all FIs and ARs within the space from the sample. In this study, techniques specialized on association detection algorithms have been examined from those developed for sample size determination [3–5, 7]. Sample size determination techniques are divided into two groups to minimize the FI and AR loss.

When the association algorithms will be run using the same parameters, support and confidence values calculated from the sample appear to be different than their counterparts calculated from the space. This margin of error is measured using two different methods. When calculating absolute margin of error, the absolute value of the difference between values from the space and the sample is considered. The relative margin of error is calculated by dividing absolute margin of error into the value within the space. In **Table 1**, the lines containing “absolute” at the “type of technique” column aim at reducing absolute margin of error, while those containing “relative” aim to minimize relative margin of error.

All examined techniques are shown in **Table 1** with suggested formulas and type of formula. The values found through the techniques determine the minimum number of transactions required for sample creation. The number of transactions which are equal to the values found is selected from the space by the preferred sample creation method.

Sample size determination techniques determine the minimum number of transactions required for sample creation. The number of transactions which are equal to the values found is selected from the space by the preferred sample creation method.

The complexity of the space is calculated theoretically using the Vapnik-Chervonenkis size [20]. Assuming that the transactions within the database are sorted according to their number of elements and that the “number of transactions” and “number of elements” are plotted on the coordinate system, the d-index value would correspond to the edge length of the largest square.

Description	Type of Technique	Formula
Zaki	FI-absolute	$\frac{-2 \ln (1-\gamma)}{\Theta \delta^2}$
Toivonen	FI-absolute	$\frac{1}{2 \varepsilon^2} \ln \frac{2}{\delta}$
Chakaravarthy	FI-absolute	$\frac{24}{(1-\varepsilon) \varepsilon^2 \Theta} \left(\Delta + 5 + \ln \frac{4}{(1-\varepsilon) \Theta \delta} \right)$
Chakaravarthy	AR-absolute	$\frac{48}{(1-\varepsilon) \varepsilon^2 \Theta} \left(\Delta + 5 + \ln \frac{5}{(1-\varepsilon) \Theta \delta} \right)$
Riondato	FI-absolute	$\frac{4c}{\varepsilon^2} \left(v + \ln \frac{1}{\delta} \right)$
Riondato	FI-relative	$\frac{4(2+\varepsilon)c}{\varepsilon^2(2-\varepsilon)\Theta} \left(v \ln \frac{2+\varepsilon}{\Theta(2-\varepsilon)} + \ln \frac{1}{\delta} \right)$
Riondato	AR-absolute	$\frac{c}{\eta^2 p} \left(v \ln \frac{1}{p} + \ln \frac{1}{\delta} \right)$
Riondato	AR-relative	$\frac{c}{\eta^2 p} \left(v \ln \frac{1}{p} + \ln \frac{1}{\delta} \right)$

Minimal sample size can be determined in terms of accuracy ε , probability of error δ , minimum support Θ , minimum confidence γ , d-index value of the space v , maximal process length of the space Δ , and the constant c . In formulas, the value η is calculated depending on variables Θ , γ , and ε ; and the value p is calculated depending on values η and Θ .

Table 1.
Sample size calculation techniques are provided.

In this study, we use a d-index algorithm, which does not seek a sorting requirement among the transactions, and it calculates v (d-index value), by initializing with 1 and by increasing it. All transactions within the database should be scanned to find the value. Where a number of transactions are large and the item number within each transaction is less, such as banking data, the length of transactions becomes decisive in determining the d-index value. In d-index algorithm, the transactions within the database were sorted in descending order of item numbers, and the value v was calculated decreasingly beginning from the maximal transaction length. Here, it is not necessary to scan all transactions.

5. Experimental evaluation and results

The tests were performed on product ownership data of banking customers. Statistical studies' code development was performed on the widely used R programming language.

When tests were performed, the steps below were followed:

1. Determine the sample size utilizing various techniques
2. Create three different samples for each technique using the simple random sampling method
3. Compare the representability of the space for the obtained sample examination with χ^2 and K-S tests
4. Use the Apriori algorithm included in the *arules* package of R language, and determine the FI and AR through the space and sample
5. Calculate the absolute error in support and trust values, and compare the results with those obtained from the space
6. Compare the duration of obtaining AR and the duration of sample creation. Generate AR from the sample

Theoretically, it is expected that the samples in various sizes obtained from FI and AR results are tuned with the results from the space, that there is a correspondence between representability and absolute error, and that the duration of transactions made on the sample and the memory consumption reduce.

To accelerate the test processes, instead of 143 products of the bank, 10 different product groups were determined, and the association between those groups was examined. The utilized banking data is a matrix including 1,048,575 customers and an ownership status of customers about 10 different product groups. The lines represent customers and the columns represent product groups. In case the customer owns a product, the intersection of that line-column indicates 1, otherwise 0. In these tests the following parameters were used: accuracy $\varepsilon = 0.04$, probability of error $\delta = 0.07$, minimal support value $\Theta = 0.02$, and minimum confidence $\gamma = 0.06, 0.1, \text{ and } 0.14$.

Table 2 shows varying sample sizes corresponding to varying minimum confidence values. Because γ was not used as a parameter in formulas Toivonen, Chakaravarthy FI-absolute, Chakaravarthy AR-absolute, Riondato FI-absolute, and Riondato FI-relative, there are no variations in calculated sizes.

When **Table 2** was examined in detail, it is obvious that the sizes obtained from the techniques Chakaravarthy FI-absolute, Chakaravarthy AR-absolute, Riondato

Description	Type of Technique	$\gamma = 0.06$	$\gamma = 0.1$	$\gamma = 0.14$
Zaki	FI-absolute	3867	6585	9426
Toivonen	FI-absolute	1047	1047	1047
Chakaravarthy	FI-absolute	14842499	14842499	14842499
Chakaravarthy	AR-absolute	30033660	30033660	30033660
Riondato	FI-absolute	9574	9574	9574
Riondato	FI-relative	1458404	1458404	1458404
Riondato	AR-absolute	15057	47005	96859
Riondato	AR-relative	5468750	5468750	5468750

Techniques where the calculated size is larger than the space were not used at the sample creation step.

Table 2.
Calculated sample sizes based on varying minimum trust values are provided.

FI-relative, and Riondato AR-relative are larger than the space (1,048,575). As the aim was to reduce the data set, these techniques were not examined in the following tests. In order to minimize the error due to simple random sampling method, three samples were created for each of the Zaki, Toivonen, Riondato FI-absolute, and Riondato AR-absolute techniques.

Table 3 shows average p values calculated from χ^2 and K-S tests. As the similarity significance between the space and sample was accepted as 95%, it is expected that p values are higher than 0.05. The results indicate that values were obtained to prove an adequate statistical similarity between the space and all obtained samples. An instability is obvious regarding p values of Toivonen where the sample size does not vary. We consider this instability results from the small sample size provided by this technique.

FIs and their corresponding ARs were determined from the samples created using Apriori algorithm. To measure the similarities of FIs and ARs, absolute error was calculated through support and trust values. Zaki and Toivonen techniques were inadequate to determine all FIs and ARs existing in the space for the value $\gamma = 0.1$. Because a loss of rule was undesirable, we have observed that these two techniques were not suitable to sample creation and time consumption tests were not examined. By calculating the error by substituting incomplete values with 0, the results on **Table 4** were obtained. As expected, where absolute support error was high, an also high-confidence error was found.

In a comparative review of **Tables 3** and **4**, no relation was detected between support and confidence errors by the results obtained from χ^2 and K-S tests. We have noticed that traditional statistical measurements were inadequate in measuring the representability of the sample which was created for association mining.

In **Table 5**, durations until creation of AR are provided for the space and created samples. While the duration from sample creation until obtaining the AR was provided for the space, the time required for sample size determination, total average time required for sample creation, and obtaining the AR by simple random

Description	Type of Technique	$\gamma = 0.06$		$\gamma = 0.1$		$\gamma = 0.14$	
		χ^2	K-S	χ^2	K-S	χ^2	K-S
Zaki	FI-absolute	0.824	0.591	0.630	0.439	0.190	0.382
Toivonen	FI-absolute	0.379	0.512	0.435	0.395	0.341	0.675
Riondato	FI-absolute	0.142	0.182	0.595	0.434	0.234	0.081
Riondato	AR-absolute	0.690	0.618	0.663	0.300	0.385	0.396

All the techniques were found to be similar to the space.

Table 3.
 p values calculated from χ^2 and K-S tests were provided based on minimum trust values.

Description	Type of Technique	$\gamma = 0.06$		$\gamma = 0.1$		$\gamma = 0.14$	
		Supp.	Conf.	Supp.	Conf.	Supp.	Conf.
Zaki	FI-absolute	0.002	0.009	0.003	0.06	0.002	0.001
Toivonen	FI-absolute	0.005	0.022	0.006	0.042	0.004	0.001
Riondato	FI-absolute	0.023	0.008	0.002	0.011	0.002	0.001
Riondato	AR-absolute	0.011	0.004	0.001	0.001	0.001	0.001

Techniques where AR loss was experienced were not tested in terms of running time.

Table 4.
Average support and trust absolute error generated based on varying minimum trust values are provided.

Description	Type of Technique	$\gamma = 0.06$	$\gamma = 0.1$	$\gamma = 0.14$
Space	–	1,832	1,825	1.87
Riondato	FI-absolute	0,186	0,193	0,193
Riondato	AR-absolute	0,193	0,253	0,343

Techniques where AR loss was experienced were not tested in terms of running time.

Table 5.
The time until rule mining based on varying minimum trust values γ was given in seconds.

sampling method were provided for the samples. As expected, the time performance of all techniques for each value γ was found better than the space. Even more benefits are expected by using actually 143 different products instead of 10 product groups in these tests.

FI and AR results of the samples were compared to those generated from the space. Within the space, the mostly encountered depository (D) product has a ratio of 94%, credit card (CC) product has 11%, and installment loan (IL) product has 8%. These products are found together within the space by a ratio of 2.3%. In spite of this low support value, three different rules with high confidence were derived. “CC, IL \rightarrow D” rule was derived, and eventually, it was observed that 92% of people who have bought CC and IL also have bought D. According to another derived rule “IL, D \rightarrow CC,” it was discovered that 31% of people who bought CC and D also bought IL. It was also found through another rule CC, D \rightarrow IL that 28% of people who bought IL and D also bought CC. These rules could not be derived from the techniques Zaki and Toivonen, and information loss was experienced. Therefore, these techniques are not suitable on sample creation for association mining. As expected, when sample size increased, the obtained absolute error in results decreased.

The utilized banking data contains information about customers and the product groups of their owned products. In other words, instead of products owned by customers, banking product groups were used. This was preferred to reduce the data set sparsity and to accelerate the test speeds. So, it was observed that even tests on the space did not take longer than 2 seconds. Whenever the advantages in terms of duration seem to be in the order of seconds, tests to be conducted with a data set containing more products (or product groups) will show decisive advantages.

6. Conclusion and future work

Because AR mining process through the space takes a long time, we have aimed at determining a smaller sized sample representing the space and AR mining through that sample. For this purpose, in this book chapter, we have investigated for techniques which provide an ideal sample size specialized on association mining.

The samples were created using the simple random sampling method, and three different samples were obtained per technique. We have tried to prevent a potential noise in our results by creating multiple samples.

The similarity of samples to the space was measured by the χ^2 test and K-S test. It was obvious that after both tests the obtained values for association mining were inadequate in measuring the representability of samples. In those tests, no relationship was found among support and confidence error values. We consider that the probability of tests giving biased outputs and the inadequacy of suggested sample sizes in measuring were the reasons for having these results.

The results indicate that the duration of AR generation within the space was compared to the total time of sample size determination, sample creation, and AR generation through the sample. It was observed that each technique was better performing in terms of space results. Riondato FI-absolute and Riondato AR-absolute techniques have given good results based on calculated absolute error values. When smaller sample size and less time consumption criteria were considered, Riondato FI-absolute technique becomes favorable.


In future studies, the data set shall be renewed in this regard, and other sample methods will also be applied. Besides, results which might be related to a single data set shall be extended with tests to be performed on another data set, and the results shall be cross-checked.

Author details

Oguz Celik, Muruvvet Hasanbasoglu, Mehmet S. Aktas* and Oya Kalipsiz
Department of Computer Engineering, Yildiz Technical University Istanbul, Turkey

*Address all correspondence to: aktas@yildiz.edu.tr

IntechOpen

© 2020 The Author(s). Licensee IntechOpen. This chapter is distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/3.0>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. 

References

- [1] Jayasree V, Balan RVS. A review on data mining in banking sector. *American Journal of Applied Sciences*. 2013;**10**(10):1160-1165
- [2] Agrawal R, Imieliński T, Swami A. Mining association rules between sets of items in large databases. In: *SIGMOD '93, Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data*; 1993
- [3] Zaki MJ, Parthasarathy S, Li W, Ogihara M. Evaluation of sampling for data mining of association rules. In: *Proceedings 7th International Workshop on Research Issues in Data Engineering*; 1997
- [4] Chakaravarthy TV et al. Analysis of sampling techniques for association rule mining. In: *12th International Conference on Database Theory*; 2009
- [5] Klemettinen M et al. Finding interesting rules from large sets of discovered association rules. In: *CIKM '94: Proceedings of the 3rd International Conference on Information and Knowledge Management*; 1994
- [6] Toivonen H. Sampling large databases for association rules. In: *Proceedings of the 22nd VLDB Conference, Mumbai (Bombay), India*. 1996. pp. 134-145
- [7] Riondato M et al. Efficient discovery of association rules and frequent item sets through sampling with tight performance guarantees. *ECML PKDD*. 2012;25-41
- [8] Baeth MJ et al. An approach to custom privacy violation detection problems using big social provenance data. *Concurrency and Computation-Practice & Experience*. 2018;**30**(21):1-9
- [9] Baeth MJ et al. Detecting misinformation in social networks using provenance data. *Concurrency and Computation-Practice & Experience*. 2019;**31**(3):1-13
- [10] Riveni M et al. Application of provenance in social computing: A case study. *Concurrency and Computation-Practice & Experience*. 2019;**31**(3):1-13
- [11] Tas Y et al. An approach to standalone provenance systems for big provenance data. In: *The International Conference on Semantics, Knowledge and Grids on Big Data (SKG-16)*; 2016. pp. 9-16
- [12] Aktas MS. Hybrid cloud computing monitoring software architecture. *Concurrency and Computation: Practice and Experience*. 2018;**30**(21):1-9
- [13] Aktas MS et al. A web based conversational case-based recommender system for ontology aided metadata discovery. In: *The 5th IEEE/ACM International Workshop on Grid Computing*. 2004. pp. 69-75
- [14] Aktas MS et al. Fault tolerant high-performance information services for dynamic collections of Grid and Web services. *Future Generation Computer Systems*. 2007;**23**(3):317-337
- [15] Pierce ME et al. The QuakeSim project: Web services for managing geophysical data and applications. *Pure and Applied Geophysics*. 2008;**165**(3-4):635-651
- [16] Aydin G et al. SERVOnGrid complexity computational environments (CCE) integrated performance analysis. In: *The 6th IEEE/ACM International Workshop on Grid Computing*; 2005. pp. 256-261
- [17] Celik O et al. Implementation of data preprocessing techniques on

distributed big data platforms. In: 4th International Conference on Computer Science and Engineering (UBMK); 2019. pp. 73-78

[18] Eltabakh MY et al. Incremental mining for frequent patterns in evolving time series databases. Technical report. Department of Computer Science, Purdue University; 2008

[19] Pei J, Han J, Lu H, Nishio S, Tang S, Yang D. H-Mine: Fast and space-preserving frequent pattern mining in large databases. IIE Transactions. 2007;**39**(6):593-605. DOI: 10.1080/07408170600897460

[20] Vapnik V et al. Measuring the VC-dimension of a learning machine. Neural Computation. 1994;**6**(5):851-876