Global Journal of Computer Science and Technology: C
Software & Data Engineering
Volume 16 Issue 2 Version 1.0 Year 2016
Type: Double Blind Peer Reviewed International Research Journal
Publisher: Global Journals Inc. (USA)
Online ISSN: 0975-4172 & Print ISSN: 0975-4350

A Systematic Review of Learning based Notion Change Acceptance Strategies for Incremental Mining

By D. S. S K. Dhanalakshmi & Dr. Ch.Suneetha

CMR college of Engineering & Technology, India

Abstract- The data generated contemporarily from different communication environments is dynamic in content different from the earlier static data environments. The high speed streams have huge digital data transmitted with rapid context changes unlike static environments where the data is mostly stationery. The process of extracting, classifying, and exploring relevant information from enormous flowing and high speed varying streaming data has several inapplicable issues when static data based strategies are applied. The learning strategies of static data are based on observable and established notion changes for exploring the data whereas in high speed data streams there are no fixed rules or drift strategies existing beforehand and the classification mechanisms have to develop their own learning schemes in terms of the notion changes and Notion Change Acceptance by changing the existing notion, or substituting the existing notion, or creating new notions with evaluation in the classification process in terms of the previous, existing, and the newer incoming notions. The research in this field has devised numerous data stream mining strategies for determining, predicting, and establishing the notion changes in the process of exploring and accurately predicting the next notion change occurrences in Notion Change.

Keywords: notion change, defencing notion change, conventional learning, supervised notion change acceptance, unsupervised notion change acceptance, data stream mining and concept evolution.

GJCST-C Classification : H.2.8, D.3.4, D.2.3

A Systematic Review of Learning based Notion Change Acceptance Strategies for Incremental Mining

D. S. S K. Dhanalakshmi [α] & Dr. Ch.Suneetha [σ]

Abstract- The data generated contemporarily from different communication environments is dynamic in content different from the earlier static data environments. The high speed streams have huge digital data transmitted with rapid context changes unlike static environments where the data is mostly stationery. The process of extracting, classifying, and exploring relevant information from enormous flowing and high speed varying streaming data has several inapplicable issues when static data based strategies are applied. The learning strategies of static data are based on observable and established notion changes for exploring the data whereas in high speed data streams there are no fixed rules or drift strategies existing beforehand and the classification mechanisms have to develop their own learning schemes in terms of the notion changes and Notion Change Acceptance by changing the existing notion, or substituting the existing notion, or creating new notions with evaluation in the classification process in terms of the previous, existing, and the newer incoming notions. The research in this field has devised numerous data stream mining strategies for determining, predicting, and establishing the notion changes in the process of exploring and accurately predicting the next notion change occurrences in Notion Change. In this context of feasible relevant better knowledge discovery in this paper we have given an illustration with nomenclature of various contemporarily affirmed models of benchmark in data stream mining for adapting the Notion Change.

Keywords: notion change, defencing notion change, conventional learning, supervised notion change acceptance, unsupervised notion change acceptance, data stream mining and concept evolution.

I. Introduction

The data streams generated in real time are dynamic in content unlike the contemporary static data environments and involve huge volumes of data transmitted at great speeds. These dynamic data communication environments are used in various fields such as, real time surveillance and monitoring, web traffic internet networks, applications producing huge HTTP data requests, weather monitoring or environment systems, RFID and wireless sensor networks, retail transactions, real time media streaming processes, cloud automations systems, telephone networks, etc.

The applications of data streams mining are many such as, financial analysis of stock market drifts, customer data transactions analysis, predicting customer preferences in retail or online shopping, telephone call records analysis, fraud prevention, social networks user content generation, internet networks traffic mining for knowledge exploration, in spam and intrusion detection, etc. The data generated by the World Wide Web in the year 2013 is stated to be around 4 zettabytes of data [1] and this is growing with magnified volumes and speed continuously. In this context and the many application areas of mining streaming data the research of real world data classification has acquired great importance both for researchers as well as for the business community.

The incremental mining process has to be associated with an efficient strategy to handle the huge volume of dynamically changing data streams that do not have any established notions in Notion Change [2] [3] and the learning algorithm applied over the drifting data streams has to find context changes in terms of the drift. The process of Notion Change Acceptance in data streams involves the objects of the streaming data categorized in terms of a concept individually either as positive or as negative case concepts. The newer concepts are mined and analyzed with visible case concepts variables based on recent topology information with a learning algorithm for predicting each incoming streaming objects case concept. The learning process strategy of adapting to the notion change is either with a set of fixed notions called supervised learning, or uses the dynamic notion which is called as unsupervised learning strategy. The learning model in exploration and extraction of the information incorporates successful predictions and if the prediction is in contrast to the objects real case concept, it changes the existing model with deletion of the old concepts.

The research in recent years has generated several strategies of benchmark Notion Change Acceptance. To further widen the scope of understanding and aiding future studies in this field we in this paper review existing strategies available in recent literature in terms of their, advantages, weaknesses, applicability, and compatibility with various domain streams and the wide scope of Notion Change

Author α : Assistant Professor, Dept., of MCA, CMR college of Engineering & Technology, Hyderabad-501401, Telangana State.
e-mail : dasaridhana123@gmail.com
Author ρ : Associate Professor, Dept., of Computer Applications, RVR & JC College of Engineering, Chowdavaram-522019, Andhra Pradesh.
e-mail : suneethachittineni@gmail.com

influenced streaming data under different contexts of notion change. Also in this paper we review the notion progress in the process of adapting Notion Change and a description of the nomenclature of data stream mining.

The remaining sections of the paper is structured as follows; Section 2 assesses the streaming data based mining nomenclature and the Notion Change impact; Section 3 reviews the models of benchmark devised and affirmed in contemporary literature. Section 4 gives conclusion of this paper with future research scope.

II. The Streaming data and Notion Change Acceptance

a) *Incremental mining*

The data streams mining process involves processing, classifying the dynamic data where the concepts might change, appear or not appear again requiring constant adaptation according to the notion change and in accordance to the data influx speed for an efficient exploration and retrieval of hidden relevant contexts.

b) *Notion Change and the objective*

The objective of learning over streaming data must be for noticing the change of notion or Notion Change by applying efficient mining strategies with the learning mechanism. If we consider the content streaming sites a considerable notion change of viewer's preferences alters the data streams in terms of the drift of concepts. Hence the principal aim in learning and mining of user preferences must be to recognize the notions changes.

c) *Notion Change Acceptance*

The notion change due to Notion Change based on a learning mechanism is adapted with appropriate learning model changes for mining efficiency and for significant data retrieval.

i. *Notion changeover frequency*

The frequency of notion change or "speed of the Notion Change" in streaming data is the average recorded time for every Notion Change Acceptance occurrence. The learning mechanism of drift Acceptance is of 2 different kinds, regular Notion Change, and impulsive Notion Change. In regular Notion Change the Notion Change event is reflected in fixed time intervals for the probable prediction of the time of next Notion Change occurrence which is usually a recursive Notion Change that converts the data to earlier state. The impulsive Notion Change learning mechanism tries to replicate unexpected and irregular Notion Change occurrences.

d) *The taxonomy of Notion Change Acceptance strategies*

The strategies for adapting the Notion Change implement the learning mechanism in two stages; stage 1 involves determining the significant notion change which has importance towards the Notion Change, and stage 2 adapts the data streams newest state into the learning process. The Notion Change adaptation into the existing learning mechanism may be ordered and explained as below,

e) *Mutable learners*

An easy technique for Notion Changes Acceptance into the learning mechanism are the mutable learners which use supervised or semi supervised learning approaches for the data streams visible scope to be dynamically expanded or restricted in terms of the newer state of the data streamed for updating the learning model with deletion of obsolete data instances.

i. *Mutable Training set based Learners*

A type of Notion Change class which use a mutable training set in the learning process these learners use a strategy of unsupervised learning based on a window of a set records grouped by considering comparable notion, or comparable instance weights. The classifiers based on the windows newest notion or newest instances having comparable weights mute the existing training set and update the learning set applied over the data streams for Notion Change Acceptance.

ii. *Colle ctive Learners*

A strategy that applies multiple learners' collectively is a well known standard data engineering approach for its realistically achievable efficiency. A learning mechanism based on diverse classifiers may be applied over streaming data with either unassigned notion changes or imbalanced data classes for achieving substantial variation in the learning process. The collective learner strategy may be applied over identical data to expand the achievable accuracy with the inclusion of a predictive classifier. A learning mechanism not based on multiple learners' experiences over fitting and decrease of performance.

Another extensively important concept for leaning over Notion Changes prone data streams in the progression of new classes. This notion progression maybe further defined using two types of networks, internet networks using the learning mechanism for intrusion prevention, and social networks use for identifying initiation of new trends. In Internet networks the notion progression is visible in case we associate a class label to every kind of attack and when the traffic is under an entirely new type of attack it results in notion progression. The social networks data streams have class labels associated with trends and the origin of a new trend whose posts are unlike the earlier posts enables notion progression. However very less importance has been given to notion progression in the contemporary research.

III. Notion Change Acceptance Strategies

The decision trees is a type mutable learner's model C4.5 [4] to be specific. The model Very Fast Decision Tree (VFDT) [5] is known as one of the initial models designed based on decision trees. The essentials of decision-tree learning are used in the design of the VFDT algorithm by which accurate and very fast decision trees are generated. These decision trees formed for data streams with the VFDT model are based on the application of the Hoeffding tree algorithm where a comparable notion subset is produced using Hoeffding bounds [6], [7]. This model mines real time data streams which are imperfect or having uncertainty characteristics depicting the entire streams as a unique advanced model in which with the arrival of newer data the decision tree's existing database is constantly updated making it more efficient in the prediction of drift in new incoming data.

The problem with this model in incremental mining is mostly due to the noise in the data used for training which forms unnecessary tree branches causing over-fitting that is further complicated because of the run-time memory inadequacy in the total decision tree accommodation declining the prediction accuracy frequently. This affects the main reason for implementing this model which is the achievable reasonable accuracy. This need for rapidly adapting the Notion Change with precision has made the model undergo several revisions.

The rule k Nearest Neighbor (kNN) is one of the earliest classification rules and also the easiest one that has been researched widely for numerous different objectives in various fields especially Notion Change Acceptance. The kNN algorithm is a type of incremental classifier which does not include any previous conventions of the data distribution and with the rapidly changing streaming data performs the learning and training continuously updating the classification model. The major difficulty with the approach however is it initiates the learning only during the time of prediction which increases the overhead in terms of time and cost especially in case of instances related to multi-label data. Also the approach is involved with computational complications when used with non-incremental type of base classifiers. A revision of the kNN based algorithm for streaming data by Alippi and Roveri [8][9] in case of a streaming data not under Notion Change is based on choosing k samples from the data using the theory based outcomes by Fukunga [10] where these newer examples are added to the knowledge base and the kNN based classifier is updated. In case of Notion Change the approach retains all the newer examples and eliminates all the old examples from its knowledge base. For data streams drifting regularly at a lesser pace a revised model is presented later by the authors, called adaptive weighted kNN which uses the strategy of assigning weights to the examples based on their nearness to the present concept where older instances comparatively still have considerably higher weights associated.

A Notion Change learning process for a streaming data is an algorithms capability of learning incrementally the newer incoming streaming information while maintaining the earlier data in the classification process. The research of Carpenter, G., Grossberg, S., Markuzon, et. al., of this difficult problem has led to the design of the Adaptive Resonance Theory (ART) model for effective classification and prediction of concept change with a model called ARTMAP (Adaptive Resonance Theory Map) [11] which is a strategy of unsupervised learning that recognizes from the data set all the different patterns incrementally with cluster formation and applies supervised learning over these clusters in the classification. The attributes of every cluster is used to map the cluster to a class considering class compatibility in terms of their labels. Carpenter, G., Grossberg, et. al., devised a Fuzzy ARTMAP [12] which applies fuzzy principals in the ARTMAP model with 2 variants of the ART model called ARTa and ARTb which are connected with an inter Adaptive Resonance module. The strategy of assessing the patterns recognized by the model is implemented with an unsupervised ART's model and the prediction process is implemented with a supervised ART model in an incremental order. In this prediction process a class in ARTa is linked to a class in ARTb and this mapped field is used to form predictive classes in learning the class associations. In case of an incompatibility scenario with existing classes the search is either repeated or a new clusters is created for properly including the newer input patterns that are dissimilar to the earlier observed examples. The incremental rule pruning strategy for fuzzy ARTMAP by Andres-Andres, A., Gomez-Sanchez, E., BoteLorenzo, M., in [13] extends the fuzzy ARTMAP models devised earlier. The model is based on updating fuzzy rules frequently with dynamic pruning the inactive and or obsolete fuzzy rules based on a pruning strategy in the paper [14] which prunes the rules set in terms of their attributes, rule confidence, rule usage frequency and rule significance.

These models of ARTMAP and Fuzzy ARTMAP are used widely in the process of incremental learning. However the problem with these models is with noisy training data where the performance becomes ineffective. The fuzzy ARTMAP model constructs maximum possible classes for learning the entire static training set and in the statistical assessment, which due to overfit leads to pending of parameter selection and the resulting generalization is ineffective. These problems are overcome with the strategy proposed in [13]. However an assessment of all the 3 models [11][12][13], considering recursive concept impacted Notion Change of high frequency shows the rule update process to be computationally complex and redundant in comparison to other similar methods.

The model AO-DCS (Attribute-oriented Dynamic Classifier) devised by Xingquan Zhu, Xindong Wu et.al., [15] is a supervised learning approach based on a single best learning algorithm whose performance efficiency is far advanced in contrast to other collective learners currently existing. This approach instead of a CC (Classifier Combination) method uses a CS (Classifier Selection) technique called DCS (Dynamic Classifier Selection) to overcome the inefficiency of the Classifier Combination method in mining highly noisy data streams under dramatic Notion Changes. The dynamic classifier selection scheme uses attribute values of instances to partition the evaluation set into subsets. This approach uses instead of the clustering technique the attribute values of the evaluation set to classify the data set into a number of small sets and the new examples are used to find the final subsets. The existing base classifiers are applied on these subsets with new examples to evaluate the performance effectiveness of a base classifier in terms of a specific domain and determine the choice of a best classifier.

The experiments are executed with 8 datasets of benchmark data streams of the UCI database repository comprising of synthetic and real time data. The experiments using a real time simulated scenario evaluates the systems performance for incremental mining under dramatic Notion Changes applying different DCS approaches with different factors like scalability, robustness and accuracy. In this process prior to data partitioning several levels of class noise or manual errors are fed into the data stream. The execution of every experiment is assessed with 10-fold cross-validation and the obtained average accuracy is used as the final result. The test outcomes show an enhanced performance with the devised DCS method compared to most other CC or CS based approaches like SAM, CVM, DCS_LA and Referee in mining real-time data streams. However a major problem with this approach is in case of high frequency Notion Changes and since the accuracy is inversely proportional to the Notion Change frequency, the learning factors of accuracy, scalability and robustness decrease the performance.

A Notion Change Rule mining Tree (CDR-Tree), for exploring, finding and precisely assessing Notion Change rules by Chien-I Lee, Cheng-Jung Tsai, Jhe-Hao Wu, and Wei-Pang Yang [16] is a unique and different approach of determining the Notion Change causes. The previous approaches devised were based on the strategy of modifying the current database for accurately classifying the incoming data and not for finding the reasons for the drift occurrences. The authors of the CDR-Tree model represent the reasons of drift occurrences in terms of categorically ordered rules set based on which the examples of old data and the incoming newer data related to different time periods are coupled to create a CDR-Tree where IG (information Gain) is used to find the node's split point in the process of forming the CDR-Tree structure. The defined Notion Change rules set is further defined by CDR-Tree with RS (Rule Support) and a RC (Rule Confidence) to screen less important instances with user specific threshold values that can be set for notable rules.

The experiments were performed with Microsoft VC++ 6.0 to depict the CDR-Tree and IBM Data Generator is used for the generating the experimental data comprising of 1 Boolean target class and 9 basic attributes given by 4 random classification functions. Here 20 integrated data sets with 6 dissimilar drift levels are tested and the results show the proposed approach achieves high accuracy in all the 20 data sets considered. The devised approach overcomes the limitations of the earlier strategies which are unable to continue the node split process in case of real time streaming data. The model is able to correctly compute the drift in case of data streams truly under Notion Change. However the concept-drifting rules in case of higher cNotion Change levels make the CDR-Tree more complex affecting the accuracy of mining. To reduce the complexity of the CDR-Tree discretization algorithm are proposed which however fail in achieving the desired accuracy. Also the chances of tree construction are highly reduced in case of streaming data under recursive concepts based Notion Change.

A stacking style ensemble-based strategy by Yang Zhang, Xue Li, [17] is devised to address the problem of single class classification of Notion Change influenced high speed and constantly changing noisy data streams with limited memory space. The objective is to use few class labels during training. A stacking strategy based ensemble learning approach is used for classifying the Notion Change exposed texts. The approach presumes the data to be coming in batches streamed in varying lengths. The classification uses only a single class and every batch streamed is classified using very few training samples. The training data selected from every batch is a positive training data set of k number of documents selected initially from 2 scenarios with the remaining data used as unlabeled data. For a more reliable sample extraction subsequently the negative samples are included along with positive samples in the data used in training. The ensemble of classifiers created from the different batches of streaming data are used to find class labels of new incoming data and inefficient classifiers in the classification process are removed from the ensemble to control the ensemble capacity dependent on limitations of memory. The algorithms devised are used to build the classifier where linear SVM classifier is used as base classifier. In the training of the base classifiers every batch is trained on 2 base classifiers, one with positive samples and another including positive samples on earlier batches as well. The learning mechanism based on ensemble stacking uses the concept descriptions preserved in its database in the learning with prediction for voting and selecting the best base classifier.

The experiments of this approach are implemented in Java simulated in WEKA with 1G memory with a dataset comprising of 20 newsgroup classes where each class has 1000 texts documents. The documents after preprocessing are vector represented with weights using the TFIDF algorithm. The simulations are done with 15 different scenarios where each scenario has 10 batches of text data. Each batch has 100 documents from each of the 20 different classes equaling to 2000 text documents. The simulations test outcomes show that the classifier achieves good performance in classification and predicting the different types of Notion Changes occurring in every batch due to variations in the user interests and distribution of data. The stacking approach is a successfully strategy for managing data streams with recursive concepts based Notion Changes. The classification efficiency achieved is higher with the devised EN methodology compared to similar window-based methods like single window (SW), fixed window (FW), and full memory (FM). However the problems with this approach are, in case of a high frequency of Notion Change the approach is unable to regulate the usage of memory where more number of stacks are required, and in case of noisy streaming data the complications associated with the process also increase.

A collective learner approach by Stephen H. Bach, Marcus A. Maloof in [18] adapts a learner pair for streaming data classification with better performance compared to other contemporary approaches. In an online learning task the Notion Change learners have to be reactive and stable for detecting the frequently occurring concept changes and this aspect is used in the devised PL-NB approach's learning mechanism where a stable learner is paired with a reactive learner in the process of finding the Notion Change and securing the newly incoming target concept. The approach focuses on the most recent time period during which concept change has occurred in the streaming data. In this window of concept change the reactive learner has better accuracy for determining the Notion Change occurrence compared to the stable online learner which has better accuracy over the reactive learner in acquiring the target concept. The approach compares the performance of the two learners in a data stream under concept change occurrences for updating the existing stable learner based learning model with the newer instances gained from the reactive learner. The better performance of the reactive learner over the stable learner in predicting the Notion Change is because the stable learner strategy is based on using all the information learned in the classification process, while the reactive learner predicts considering only the information learned in training over a recent window of time during which the concept change occurs.

The simulations experiments with WEKA of the proposed PL-NB algorithm is done by combining the paired learner with the base learner using the naïve Bayes online algorithm. The execution is done with 2 variations of the PL-NB algorithm using a similar online NB algorithm as stable learner and with dissimilar reactive learners. The scheme is assessed by comparing it with 4 different schemes, NB (single base learner), DWM (dynamic weighted majority), AWE (accuracy weighted ensemble), and streaming ensemble algorithm (SEA) with 2 synthetic problem concepts, the Stagger concepts and the SEA concepts, and with 3 data sets of real time, a meeting scheduling data set, a electricity prediction data set, and a malware detection data set. The tests outcomes indicate for the above problems, the approach of paired learners has an equivalent or an enhanced performance over other schemes as it uses only an ensemble of 2 learners where the other methods use an ensemble with a higher number of learners. The approach uses lesser space, time, and cost in contrast to the high overhead incurred with the other schemes. The efficiency achieved in mining unfamiliar type of class labels with paired learner classifier is also comparatively very high. However the problem with the devised paired learner scheme is that both the classifiers are inefficient in tracking the noise in the data streams which affects their accuracy of predicting the Notion Change.

A unique framework of an ensemble classifier called WEAP-I by Zhenzheng Ouyang, Min Zhou, et. al., [19] is an approach developed based on the collective learning strategy. This design strategy combines the models of WE [31] and AP [32] for addressing the existing PL-NB approach constraints [18] in the enhancement of the performance of classifying noisy data streams. The averaging ensemble classifier AE has lesser probable occurrences of errors comparatively though in classification the accuracy is low as it is not based on future instances led alterations and evolution of concept in noisy data, and has a low stability as in training it doesn't consider older data portions. The model weighted ensemble classifier WE is capable of handling noisy data though incapable of handling concept evolution constantly. These two issues in incremental mining are effectively handled by the WEAP-1 devised by integrating the structure of an online learner WE trained on the highest possible portions of data with a reactive learner AE trained on the most recently available portion of data. In the completion of this process all the base classifiers selected are joined to create the WEAP-I ensemble classifier.

The experiments are executed in Weka with real time instruction detection data set KDDCUP'99 comprising of a series of TCP connection records which are of 2 types, one a normal connection, and the second is an instruction connection of 4 different attack types, DOS, U2R, R2L, and Probing. The tests are executed with 100 data portions where each portion has 2000 sample data, first with a normal connection and second with an attack connection where the data is not replaced in between them. Next noise is added to approximately 30% of the data and the performance evaluated and then the tests are repeated by adding noise to each selected data portion. The basic classification algorithms DT (Weka J4.8 implementation) and SVM (Weka SMO implementation) are applied over these data sets and evaluated with the parameters of classified Algorithm L, Average Accuracy (Aacc), Average Ranking (AR) and Standard Deviation. The results of the WEAP-I model shows it is more robust and efficient in solving the learning and classification problems of real time data stream irrespective of the levels of noise in training data compared to the performance of averaging probability ensemble. The difficulty associated with this model is its inefficiency in the classification considering the context of the Notion Change and its incapability of handling recursive concepts based Notion Change.

A unique E-Tree Indexing structure by Peng Zhang, Chuan Zhou et. al., in [20] is a collective learner based ensemble classifier. The approach is devised for handling cost and time impaired high speed real time data streams where the incurred constrains related overhead including process complexity increases with the data dimensionality. These problems deter a feasible ensemble learning and mining classification solution to be devised in terms of response time and overall performance efficiency. This distinct ensemble-tree or simply E-tree solution models or indexes the base classifiers to form an ensemble in an orderly way for fast decision making in the predictive process of classifying the newer instances with minimal complexity associated with the factors of time and related overhead. The strategy of this E-trees approach considers an ensemble of base classifiers as spatial databases by modeling every base classifier as a set of spatial data objects. The ensemble model E-tree is mapped to the spatial database that creates a spatial index supporting the search process of the spatial database and thus the predictive complexity associated with the new instance classification is effectively minimized. In this classification approach the E-tree is searched for every new instance and from the leaf node(s) the decision rules related to the new instance are determined and merged for predicting its class label. A new classifier thus formed is merged with the E-tree structure and a new entry associated to this new classifier is created in the database and the retrieved decision rules are sequentially inserted and further connected in the tree structure. The classifiers that are old and inapplicable in terms of the newer instances in the classification due to overcapacity are removed from the E-tree ensemble which might otherwise lead to increase in the process cost. The E-trees ensemble model evolves with constant and automatic updating process which adds the incoming new classifiers and deletes the old inapplicable classifiers and adapts to the streaming data's latest patterns and trends. The E-trees are devised for binary classification only whereas to a certain extent the multi-class problems are solved with an E-forests model that merges several E-trees.

The experiments for assessed the E-trees performance is done in terms of prediction time, memory usage, and prediction accuracy with 3 real-time and synthetic data streams intrusion detection, spam detection, malicious url detection collected from the UCI repository. F-Score is used for feature selection and the devised approach is compared with 4 benchmark models Global E-tree (GE-tree), Local E-tree (LE-tree), Global Ensemble (G-Ensemble), and Local Ensemble (L-Ensemble) where the decision trees algorithms C4.5 is used for training and retrieving the data rules. The assessment of the online query traversal in the devised E-tree methods is analyzed and compared with 4 methods, the TS model, the fractal model, selectivity method, and the ERF model and is done with 3 measures, time-cost, memory cost, and accuracy with a decision rules set of total 200 rules used to quantify the average relative error. These benchmark approaches are compared with each other with varied ensemble size, node size and target indexing class and 10 data sets. The performance of our approach demonstrate that LE-tree outperforms all other methods, is faster with lesser prediction time, and occupies lesser memory with the exception compared to L-Ensemble approach where the proposed approach consumes more memory significantly. The method effectively contributes towards achieving accuracy of prediction comparatively and the approach may also be implemented with different other types of classifications not related to ensemble learning and for data analysis of spatial or temporal databases also. The model does not effectively describe the Notion Change supervision and prediction and lacks proper assessment of Notion Change and of the class labels temporal validity.

An approach devised for solving the data stream classification problems is proposed in the paper [21] by Kapil Wankhade, Snehlata Dongre, et,al, is a supervised learning based strategy devised for achieving high accuracy in the classification performance of high speed, huge size, and noisy Notion Change influenced data streams. The devised models strategy is based on using two different methods the weighted majority method together with the method of adaptive sliding window for achieving the objective of achieving better and high classification accuracy over other models. In this model the approach polls a new example by all the ensembles algorithms considered experts. The predictions polled and the weights linked to the algorithms are combined, and in terms of the maximum accumulated weights it determines the global prediction of the labels of the class. The prediction accuracy is improved by incremental learning where incorrect predications by an expert has the related algorithms weight being reduced and the process repeated where experts with below the threshold values are deleted and new experts created. The performance is further improved by normalization of the weights where each expert is scaled according to the maximum weight so that the decision and prediction process is not totally influenced by the recently created experts. The weighted majority technique thus accurately classifies the Notion Changeing data streams mostly with noise. The accuracy in processing the fast streaming data is achieved with the sliding window concept which monitors the existing learning model and if the pace of change is greater than a set threshold value the windows obsolete sections are automatically removed from the strategy and the model gets updated by the base learners according to statistically determined distribution changes. This learning and classification is very fast in pace with the speed of the Notion Changing streaming data using sub linear memory

The experiments are performed with existing models Oz a Bag, Oz a Boost, OC Boost, Oz a Bag ADWIN, AEBC, and the devised model. The datasets used in the experiments are synthetic datasets of two types' hype plane and RBF where the Notion Change is synthetically applied and with real datasets of the UCIML repository. These approaches are tested with factors of accuracy, time, and memory. The devised model aims for better accuracy so in terms of classification accuracy it shows performance improvement compared to the other models.

The study by G. R. Marrs • M. M. Black et.al., of the streaming and Notion Change influenced data classification devise an approach [22] based on the latency of new instances arriving and the importance of the time stamp of the instances in the life cycle of the learning process. The authors apply a time stamp based learning strategy with latency applied arbitrarily on the data resulting in new rules of classification. The proposed model has 2 algorithms CDTC 1 and CDTC 2 which use the time stamp protocol or time of classification protocol for a latency impacted data classification with a proper definition given for the ordering of the instances selected in a temporal environment.

The experiments with 4 online learners, the contemporary CD3 and CD5 algorithms and the time stamp based proposed meta data tagging protocol approaches CDTC version 1 and CDTC version 2 are implemented with different scenarios of latency based Notion Change influenced streaming data. The tests with a normal latency shows, the CDTC algorithms ver 1 and ver 2 are immediately affected by the drift and the recovery is much faster and the rate of classification achieved is much greater before occurrence of another drift. The approach shows equivalent performance with other domains such as binary class value, airplane arrival data and real protein data which validates the time stamp protocol performance overcoming the constraints of memory and time for different classification scenarios.

A new approach for data stream classification devised by Zohre Karimi, Hassan Abolhassani et. al·[23] handles batch data with discrete and continuous variables, the data streams of huge volume for reduced overhead incurrence. The devised approach is a batch classifier based on the harmony search algorithm called harmony-based classifier (HC) in which the every classifier is a potential solution determined by user specified parameter based rules for the selection of a class. A Harmony is defined by the user parameters set in terms of variables sourced from memory which can be changed as per user requirements and the fitness of a harmony is determined by its accuracy. The performance of an incoming classifier if is efficient compared to a least performing classifier in the memory it is substituted and the obtained classification model is used for class label prediction. The HC approach is not capable for handling streaming data where there is no pre-determined training data available and so is combined with the Stream Miner framework for a new classification model called IHC (Incremental Harmony-based Classifier). The evaluation of the fitness by the IHC is done by a detecting and incrementally learning mechanism over the Notion Change influenced data streams with n-time cross validation towards determining the classifiers accuracy and selecting the final classifier with maximum accuracy. The IHC approach is further improved for the method called IIHC (Improved incremental harmony-based classifier) for handling the overhead incurred due to computation of learning stable and recurring concepts and learning data with noise for increasing the robustness of the model.

The experiments of the IIHC model are performed with 8 benchmark data sets of real world and synthetic datasets known for their accurateness in prediction The outcomes of the performed experiments prove that compared to other classifiers available for streaming data classification the speed and accuracy achieved with the IIHC classifier is improved for predicting the drift and is also robust in performance in data impacted by noise. However the issues of lesser important Notion Change and the recursive concepts based Notion Change are not properly assessed.

An approach devised by Mayank Pal Singh in the paper [24] is a novel approach that uses a strategy of supervised adaptive learning with fixed window that identifies the Notion Change, trains, updates, and evolves the model continually in the classification process of the data. The devised model performs data classification using a classifier based on the Naïve Bayes theorem. The incoming traffic is separated into ingress and egress traffic and the related attributes like Source IP, Destination IP, Source Port, Destination Port, Flags, Protocol are extracted. The training dataset termed as base class is used to classify the current class data set collected from the incoming streaming data. The examples of the base class are linked to the current class examples using the NB classifier and the resulting ROC curves is used to determine and quantify the Notion Change occurring. The devised model finds the drift occurrence using the ROC curve and identifies the flow specific data attributes responsible for the drifting concept.

The experiments are performed with the WEKA simulation tool on lab collected real time dataset and on the KDD datasets. The classification is implemented with the complete dataset and also using the flow specific attributes with a training window ranging from a few hours to a couple of days depending on the data under drift. The drift is generated in the traffic by using a packet generator tool that injects in normal traffic a protocol based traffic which causes drift to occur. The analysis of the results show for a KDD dataset the model is able to correctly distinguish normal and anomalous traffic. The model may be used with other classifiers as a pre-processing tool for better classification. The models classification performance in terms of the cost incurred and the accuracy achieved may be further enhanced. However the model does not totally validate the importance of data streams characterized by capricious data.

An unsupervised clustering framework that is an on-demand resources aware classification strategy defined by conditional rules called SRASTREAM is proposed in the paper [25] by Gansen Zhao, Ziliu Li, Fujiao Liu, et.al,. The methodologies available now focus on the accuracy or on the speed whereas the devised approach based on the resource available classifies the data streams. If there is no drifting of the concept the approach does not perform the clustering and if the Notion Change occurs then the cluster refining is done in terms of the drift detected which greatly reduces the time and cost overhead and makes possible the mining of huge streaming data in real-time. The devised framework combines different tasks such as clustering, computing, evolution detection and resource monitoring.

The experiments performed are 3 comparison tests with the devised approach and existing approach CluStream. The datasets used are the KDDCUP99 data and synthetic dataset. The results of the tests show clustering performance with the proposed approach is capable of specifically clustering data of huge data size. The proposed results of the approach do not specifically validate the approach and the model is unable to completely address the issue of recursive concept based Notion Change.

A new ensemble classifier called Rot-SiLA by Muhammad Shaheryar, Mehrosh Khalid and Ali Mustafa Qamar [26] is a collective learner approach which has Rotation Forest algorithm [30] integrated with the Similarity Learning Algorithm (SiLA) ([29]. The classification strategy of the approach is devised based on similarity where relevant similarity metrics are used instead of the distance measure. The Rotation Forest classifier can be used with different selections of base classifiers and is a feature extraction based strategy which uses the PCA (Principal Component Analysis) technique to divide the feature set into K subsets and maintains all the principal components information in the process of classification.The Similarity Learning Algorithm (SiLA approach strategy is built by integrating kNN (k nearest neighbor) algorithm with Voted Perceptron technique and the learning strategy for classifying any kind of data uses the related similarity metrics instead of the distances. The assigning of an example by the Rot-SiLA algorithm to a specific nearest class has the similarity associated to a class equal to the total all the similarities existing among an example being classified and all the k nearest neighbors in the class.

The experiments are done with a fourteen UCI benchmark datasets of different domains such as medical, biology, and materials classified first with SiLA using kNN-A and SkNN-A and then with the ensemble learner Rot-SiLA kNNA and Rot-SiLA SkNN-A algorithms. The learning schemes classification accuracies gained with the Rot-SiLA ensemble learners are compared with the SiLA kNNA and SiLA SkNN-A and also with the Rotation Forest ensemble which has various integrations with dissimilar base classifiers. The test outcomes show the devised models is optimal compared to the other existing approaches. However as the extracted feature set is first separated into subsets with the PCA technique the devised models accuracy is defined by the accuracy of the variance matrix formation in the principal component analysis process.

The SA-Miner strategy proposed by Chao-Wei Li , Kuen-Fang Jea in the paper [27] for incremental mining models the frequently occurring item sets by their frequency relationships with a support approximation strategy for definitively characterizing the data streams in terms of concepts. The algorithm SA-MINER collects the examples defining a concept with the support approximation strategy which generates the concepts for the document. The techniques of other types could be used for monitoring variations of the support relationships to find the new trends and for capturing gradual drifts.

This devised model is tested and evaluated with a number of experiments and performance compared with many approximate algorithmic methods such as Stream Mining, Loss-Counting, DSCA, and SWCA. The test data used in the experiments uses synthetic as well as real-life datasets with 3 type's metrics, space efficiency, time efficiency, and mining quality. The criteria of the tests performed are set as maximum or satisfactory in terms of efficiency in achieving accuracy in mining with least memory usage. The approach achieves better classification accuracy compared to the other streaming data classification strategies.

The density-based unsupervised learning approaches reviewed in the paper [28] are capable of learning data comprising of undefined cluster shapes as well with noise. This density based model for robustness and scalability combines 2 algorithms, one called micro-cluster formation algorithm and second the grid formation algorithm. The model does not use any previous clusters number information explores the Notion Changes influenced data streams. The paper reviews the important density based clustering algorithms for streaming data classification and the issues faced with these algorithms. The algorithms are classified into two type's micro-cluster and grid algorithms by the authors.

The simulation experiments of the different algorithms are done to evaluate their performance using real life data sets and with different metrics for cluster quality. The density based algorithms are able to mine data with different clusters those without any particular shape in terms of robust and scalable performance factors. However the performance of the density based algorithms is dependent on a large number of parameters and only a few algorithms are able to handle high dimensional data streams or complex clustering processes, or different other types of data streams.

IV. Conclusion

The objective of the paper is taxonomy and systematic review over incremental mining under the influence of Notion Change. The information retrieval and knowledge discovery progression from the strategies based on static data volumes has moved to the streaming data scenario where the notion change is not available, the established concept is not static but due to changes in the environment drifts with time, where the existing static data classification approaches are not applicable. The growth in the research in the data stream mining field has been propelled with the rapid developments in computing and communications where numerous organizations have varied interests in information exploration, extraction and knowledge discovery. The focus of these research activities in recent years has been for devising Notion Change Acceptance strategies for high speed and noisy data streams considering factors of higher accuracy, lower time complexity, scalability and robustness in the mining process and among these devised strategies a considerable number of them have materialized as benchmark strategies. These models of benchmark have been reviewed in this paper with their merits and demerits giving a better perception of these models for Notion Change and their algorithms for assessing their performance. The domain of research which is reviewed in this paper offers many new and superior strategies for mining streaming data under the influence of Notion Change. The research scope in this field is still huge as these existing models are not comprehensive and also not totally compatible with the many different types and domain contexts of streaming data influenced by diverse scenarios of Notion Change and notion changes. The factors like Notion Change context, temporal validity of Notion Change, and recursive concepts based Notion Change are not given the needed importance. Based on these factors the research for devising newer strategies and models for Notion Change Acceptance in data stream mining has wide opportunities. These opportunities will be the focus of our future research and design of newer models and strategies for Notion Change Acceptance.

References Références Referencias

1. https://en.wikipedia.org/wiki/Zettabyte.
2. SCHLIMMER, J. AND GRANGER, R. 1986. Incremental learning from noisy data. Mach. Learn. 1, 3, 317–354.
3. WIDMER, G. AND KUBAT, M. 1996. Learning in the presence of Notion Change and hidden contexts. Mach. Learn. 23, 1, 69–101.
4. Quinlan, J.: C4.5: programs for machine learning. Morgan Kaufmann (1993)
5. Domingos, P., Hulten, G.: Mining high-speed datastreams. In: KDD, pp. 71–80. ACM (2000)
6. Hoeffding, W.: Probability inequalities for sums of bounded random variables. JASA 58(301), 13–30 (1963)
7. Maron, O., Moore, A.W.: Hoeffding races: Accelerating model selection search for classification and function approximation. In: NIPS, pp. 59–66 (1993)
8. Alippi, C., Roveri, M.: Just-in-time adaptive classifiers in non-stationary conditions. In: IJCNN, pp. 1014–1019. IEEE (2007)
9. Alippi, C., Roveri, M.: Just-in-time adaptive classifierspart ii: Designing the classifier. TNN 19(12), 2053–2064 (2008)

10. Fukunaga, K., Hostetler, L.: Optimization of k nearest neighbor density estimates. Information Theory 19(3), 320–326 (2002)
11. Carpenter, G., Grossberg, S., Markuzon, N.,Reynolds, J., Rosen, D.: Fuzzy artmap: A neural network architecture for incremental supervised learning of analog multidimensional maps. TNN 3(5), 698–713 (1992)
12. Carpenter, G., Grossberg, S., Reynolds, J.: Artmap:Supervised real-time learning and classification of nonstationary data by a self-organizing neural network. Neural Networks 4(5), 565–588 (1991)
13. Andres-Andres, A., Gomez-Sanchez, E., BoteLorenzo, M.: Incremental rule pruning for fuzzy artmap neural network. ICANN pp. 655–660 (2005)
14. Carpenter, G., Tan, A.: Rule extraction: From neural architecture to symbolic representation. Connection Science 7(1), 3–27 (1995)
15. Xingquan Zhu, Xindong Wu, and Ying Yang, “Dynamic Classifier Selection for Effective Mining from Noisy Data Streams”, Proceedings of the Fourth IEEE International Conference on Data Mining (ICDM’04)
16. Chien-I Lee, Cheng-Jung Tsai, Jhe-Hao Wu, Wei-Pang Yang “A Decision Tree-Based Approach to Mining the Rules of Notion Change”, Fourth International Conference on Fuzzy Systems and Knowledge Discovery (FSKD 2007) IEEE.
17. Yang Zhang, Xue Li, “One-class Classification of Text Streams with Notion Change” 2008 IEEE International Conference on Data Mining Workshops.
18. Stephen H. Bach, Marcus A. Maloof, “Paired Learners for Notion Change”, Eighth IEEE International Conference on Data Mining, 2008.
19. Zhenzheng Ouyang, Min Zhou, Tao Wang, Quanyuan Wu,“Mining Concept-Drifting and Noisy Data Streams using Ensemble Classifiers”International Conference on Artificial Intelligence and Computational Intelligence, 2009.
20. Peng Zhang, Chuan Zhou, Peng Wang, Byron J. Gao, Xingquan Zhu, Li Guo, “E-Tree: An Efficient Indexing Structure for Ensemble Models on Data Streams”, IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, JUNE 2011 1.
21. Kapil Wankhade, Snehlata Dongre, Ravindra Thool, “New Evolving Ensemble Classifier for Handling Notion Changeing Data Streams”, 2nd IEEE International Conference on Parallel, Distributed and Grid Computing, pp. 657-662, 2012
22. G. R. Marrs • M. M. Black • R. J. Hickey, “The use of time stamps in handling latency and Notion Change in online learning”, Springer-Verlag, 2012.
23. Zohre Karimi · Hassan Abolhassani · Hamid Beigy, “A new method of incremental mining using harmony search”, Springer Science+Business Media, LLC February 2012.
24. Mayank Pal Singh, “Quantifying Notion Changeing in Network Traffic using ROC Curves from Naive Bayes Classifiers”, Nirma University International Conference on Engineering (NUiCONE), 2013
25. Gansen Zhao, Ziliu Li, Fujiao Liu and Yong Tang, “A Notion Changeing based Clustering Framework for Data Streams”, Fourth International Conference on Emerging Intelligent Data and Web Technologies, 2013.
26. Muhammad Shaheryar, Mehrosh Khalid and Ali Mustafa Qamar, “Rot-SiLA: A Novel Ensemble Classification approach based on Rotation Forest and Similarity Learning using Nearest Neighbor Algorithm”, 12th International Conference on Machine Learning and Applications, pp. 46-51, 2013.
27. Chao-Wei Li , Kuen-Fang Jea, “An approach of support approximation to discover frequent patterns from concept-drifting data streams based on concept learning”, Springer-Verlag London, 2013.
28. Amineh Amini, Member, IEEE, Teh Ying Wah, and Hadi Saboohi, Member, ACM, IEEE, “On Density-Based Data Streams Clustering Algorithms: A Survey”, JOURNAL OF COMPUTER SCIENCE AND TECHNOLOGY 29(1): 116–141 Jan. 2014.
29. M. Qamar, E. Gaussier, J.-P. Chevallet, and J. H. Lim, “Similarity learning for nearest neighbor classification,” in Data Mining, 2008. ICDM’08. Eighth IEEE International Conference on. IEEE, 2008, pp. 983–988.
30. J. J. Rodriguez, L. I. Kuncheva, and C. J. Alonso, “Rotation forest: A new classifier ensemble method,” Pattern Analysis and Machine Intelligence, IEEE Transactions on, vol. 28, no. 10, pp. 1619–1630, 2006.
31. M. Scholz and R. Klinkenberg. An ensemble classifier for drifting concepts. In Proc. ECML/PKDD’05 Workshop on Knowledge Discovery in Data Streams.
32. Jin Gao, Wei Fan, and Jiawei Han, On appropriate assumptions to mine data streams: Analysis and Practice, Proc. of ICDM’07, pp. 143-152