Babu and Narsimha, J Comput Sci Syst Biol 2018, 11:5
DOI: 10.4172/jcsb.1000283

Journal of
Computer Science & Systems Biology

Review Article Open Access

Survey on Dynamic Concept Drift

Kishore Babu* and PV Narsimha Rao

Department of CSE, Institute of Aeronautical Engineering, Hyderabad, India

Abstract

The role of information technology and the advancement of the cloud computing have increased the use of the data in everyday life. Data analyzation and data storage become a challenging task during the processing of this large data. In the online applications, the data stream varies with rapid speed and has a larger volume. The recurring concept drift in the data stream makes the classification process to be complicated. Various algorithms discussed in the previous research works have not effectively addressed the problem of detecting the recurring concept drift in the data. The selection of the high-performing classifier model is also a challenging research goal. This research introduces two classification models for classifying the data with the recurring concept drift in the real time environment.

Keywords: Data mining; Dynamic concept drift; Data stream; Data stream classification

Introduction

The data processing when done in the real time then the information retrieval from the data content is a more challenging task. The data mining and the machine learning allows the information retrieval from the larger data. Data mining refers to the process of extracting constructive knowledge from large databases. The various data mining techniques allow the processing of the data in the real-time applications such as cloud computing, ATM transactions, health care diagnosis, etc. The data in the real time has both the statistical and dynamic variation. Due to the large size and complexity of the data, the data mining techniques undergoes various challenges for data processing. The use of the single personal computer for the data mining of the large Data sets increases the computational cost of the data mining algorithm. In this paper, various existing Works for the data stream classification with the Concept drift were analysed. These methods help the researchers to understand the issues of the existing systems and allow producing a solution to the various problems in the data stream classification. The data Mining techniques focus more on the scalability and data access efficiency. The data mining models produce a set of results by analysing the large database. The data mining model depends on the complex mathematics or sophisticated algorithms for data processing. The data mining techniques differ for the database in a static environment and dynamic environment. The data mining observes the large database and finds the information content from the database for the efficient processing.

Data Stream Classification

The Data stream classification allows extraction of the models and the patterns from the continuous and static data environments. The data stream has a rapid rate of generation of the data from the online. The internet applications have increased the data rate and the storage of the generated data. Hence for storing and learning the nature of the data, data stream classification is necessary. The data mining involves two processes. They are clustering and classification. The efficient data stream classification solves the problems occurring in the real time environments such as real-time intrusion detection, spam filtering, and malicious website monitoring. In the dynamically varying environments, data arrive in a continuous fashion as a stream. Multiple scanning of the data stream for data classification increases the memory requirement. Compared to traditional classification of data, data stream classification faces two extra challenges such as, large/increasing data volumes for storage and drifting/evolving concepts of data stream. The main problem in the data stream classification is the concept drift. The data in the stream undergoes a shift in the concept with respect to the time. The other problem occurring in the data stream classification is concept evolution, feature evolution, limited labels and noise. The data stream classification also suffers from the class imbalance. Various approaches for the data stream classification with the concept drift are discussed here for obtaining the essential solution. Figure 1 shows the various approaches for the data stream classification with the concept drift.

Ensemble-based algorithms

Zhang et al. have proposed an Ensemble-tree (E-tree) indexing structure for the organization of the base classifiers present in an ensemble [1]. The main advantages of this E-tree are: E-trees subject for the automatic update and they treat the ensembles as spatial databases. The E-tree operation has three parts. They are Search, Insertion, and Deletion. The proposed method classifies the incoming stream by inserting the classifier into a tree and removing the outdated one. They have proposed an E-tree indexing structure for the classification of the data stream at the high-speed environment. The proposed model does the classification at the sublinear time complexity. The proposed E-tree motivates the real-time applications by efficiently addressing the prediction efficiency problem. The E-tree has advantageous techniques such as spatial indexing for converting the ensemble models into the spatial databases. The spatial databases allow classification of the data stream. The proposed E-tree implements the general classification models rather than the ensemble learning. This proposed E-tree has suffered a disadvantage when used for spatial and temporal analysis of the data stream classification. Also, it does not consider the recurring concept drift in the data stream.

***Corresponding author:** Kishore Babu, Professor, Department of CSE, Institute of Aeronautical Engineering, Hyderabad, India, Tel: 9908300815; E-mail: domalakishore@gmail.com

Received July 16, 2018; **Accepted** August 06, 2018; **Published** August 20, 2018

Citation: Babu K, Narsimha RPV (2018) Survey on Dynamic Concept Drift. J Comput Sci Syst Biol 11: 256-264. doi:10.4172/jcsb.1000283

Copyright: © 2018 Babu K, et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

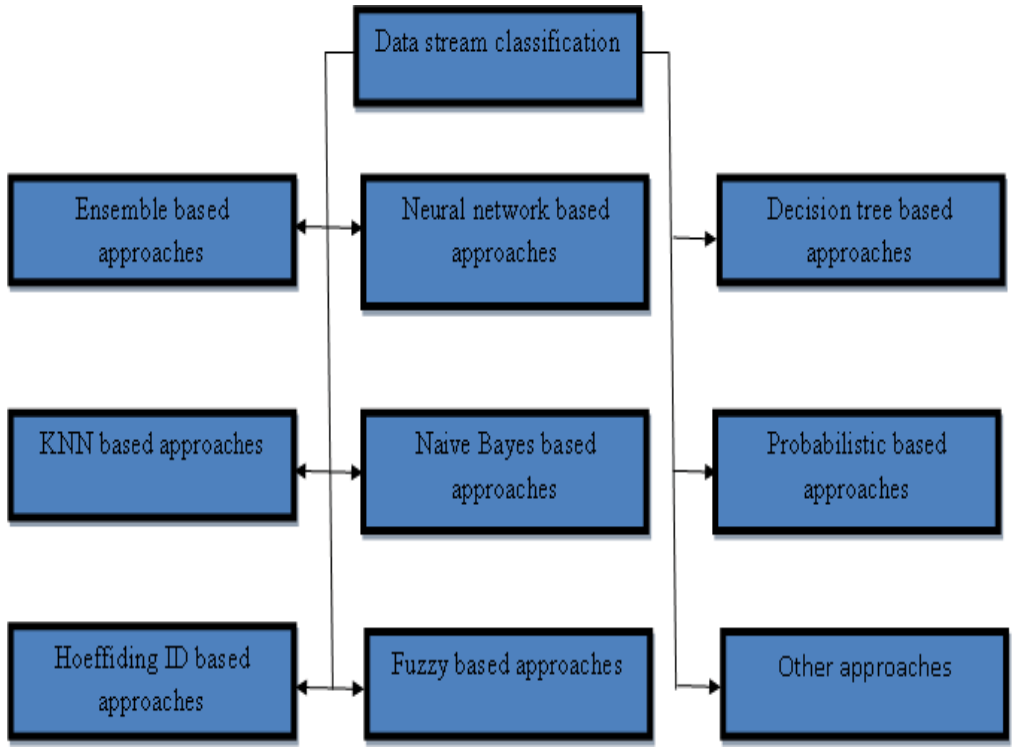

Figure 1: Data stream classification approaches.

Brzezinski and Stefanowski have proposed an Accuracy Updated Ensemble (AUE2) for the data stream classification in the real time environment [2]. This algorithm development has concentrated more for the identification of the various drifts such as recurring concept drift and concept drift in the data stream. The proposed AUE2 is the combination of the accuracy based ensemble weighing mechanism and the Hoeffding tree with the incremental nature. This algorithm development was mainly done for the block based ensemble data stream, classification. The proposed AUE2 had obtained better average classification accuracy and also reduced the usage of the memory requirement. The performances of the proposed AUE2 algorithm were evaluated based on the parameters such as classification accuracy, processing time, and memory costs. The performance results showed that the proposed model had better classification than the single classifiers. The categorization of the AUE2 algorithm falls under the block based ensemble methods since previously trained components in the AUE2 algorithm were recovered from premature reactions. The proposed AUE2 offered high classification accuracy in dynamic environments as well as in static environments with the concept drift. The memory consumption for the data stream classification was less compared to the other classifier algorithms. This proposed algorithm suffered a disadvantage during the use of the partially labeled streams.

Gama and Kosina have presented a system known as the Meta-learner for monitoring the learning process evolution [3]. This proposed model has concentrated on the field of the data learning. The data learning allowed the detection of the recurring concept drift in the incoming data stream. The meta-learners had characterized the use of the previously learned models in the technique. The proposed meta-learner system had used unlabeled and the labeled examples. The unlabelled examples helped the meta-learner to detect the recurring concept drift. Activation of the previously learned models helped the meta-learner to take pro-active actions for the data stream classification. The experimental evaluation of the proposed model had suggested the following results. The proposed meta-learner had changed the decision model for the data classification with the maximum speed and accuracy when the recurring concept drift appeared in the data stream and has provided information about the recurrence of concepts. The proposed meta-learner usage was restricted to the dynamically varying environment. When the change of concept drift occurred in the data stream, then the usage of meta-learners was found to be more effective. The proposed model allowed the reuse of the previously learned models in the classifier. The two text mining problem calculated the experimental evaluation of the proposed meta-learner. The main advantages of the meta-learners were the information update about the recurring concept drift and the quick change in the adaptation models of the classifier. The primary importance of the learning algorithm was monitoring the learning process and the ability to self-diagnosis the predictive models. The self-diagnosis had allowed the updated model to correct itself for any error change to occur in the model. The diagnosis allowed for the prediction of the failure rather than the detection of the failure. This made the proposed meta-learner algorithm to be effective for the dynamically varying environment. The disadvantage of the proposed model was the ability of reasoning and learning the classification model for various resources was less.

Raja and Swamynathan have proposed the ensemble of classifiers model for the data stream classification in ensemble database [4]. This classifier model combined two types of classifiers such as Similarity-based Data Stream Classifier (SimC) and Online Genetic Algorithm (OGA) classifier. The proposed ensemble model performance was studied through the implementation of the model in the dynamically

varying environments. The results had suggested an improved accuracy and less classification error rate for the data stream classification with concept drift. The proposed ensemble model had different real-time applications like temperature analysis and sensor data analysis in which the data stream varies abruptly. The proposed ensemble model had provided better results when compared with the existing models. The model parameter update purely depends on the accuracy enhancement of the data classification. The Similarity-based Data Stream Classifier when used along with the OGA classifier the classification of the online data can be done with ease. The use of two algorithms has increased the complexity of the classification model update. The model had predicted the concept drift in the incoming data stream and but neglects the recurring concept drift.

Haque et al. have proposed a model called HSMiner with the multitiered ensemble-based method [3]. This proposed model performed classification of the data stream with a large amount of data. Labeling the instances in the large data stream was a challenging task. In the dynamic environment involving online streaming of data has large data streams. When these data streams were subjected for the classification scaling of the large data into smaller data is a challenging task. HSMiner is a proficient method to address the challenges in data stream mining but faces a bottleneck problem. The proposed HSMiner algorithm faced a scalability issue during the classification of the data stream. This overcome with the use of the HSMiner with the three MapReduce-based approaches. The HSMiner overcome the bottleneck problem with the use of the based AdaBoost ensembles for numeric features in the data stream classification. Another disadvantage faced by the proposed model was the increase in the use of the AdaBoost ensembles when the chunk size of the data stream is larger or when the number features in the information table is larger. In this thesis, they have addressed the scalability issue by proposing the HSminer with the MapReduce-based approaches. The performance of the proposed algorithm was measured through the factors of speedup and scaleup. They had shown a significant improvement in these factors for large sized data streams.

Canzian et al. have proposed an online ensemble learning algorithm for the dynamically varying data environment [5]. This algorithm had concentrated more on the update of the aggregation rule which tends for the change when the data are dynamically varied. The probability of data classification error rate was calculated by measuring the upper bound and lower bound rate. The misclassification probability value depends on the best static aggregation rule. The probability case was found for both the best case and worst case scenario. The upper bound value tends to 0 for the worst case misclassification probability if the misclassification probability of the static aggregation rule is found to be 0. These settings of the model when applied to the dynamic environment, they provide better results. They test the performance of the proposed online ensemble learning algorithm by subjecting it to the classification of various datasets. The proposed model had also considered the network topology to analyse the flow of the data stream in online. The multi-hop communication protocol allowed the local predictions to be spread over the network even when the physical network topology is not completely connected. Their online ensemble learning algorithm had the advantages of the less communication, computational, energy, and memory storage requirements. The proposed algorithm had made use of the best static aggregation rule and the best local classifier for the data stream classification. This algorithm was further extended to reduce the difficulties faced in the dynamic environment such as asynchronous learning, receiving or not receiving the label with delay. The proposed model had shown performance gains ranging from 34% to 71% on state-of-the-art solutions for the dynamic data stream. This technique suffers a disadvantage with the use of the parallel streaming of online data.

Sethi et al. have proposed an ensemble classification approach for the data stream classification with the Spatio-Temporal drifts [6]. The proposed method does the classification even if the labeling of the data is limited in the information table. The data classification in the spatial configuration will have Spatio-Temporal drifts when the environment is changing. The proposed model identifies the Spatio-Temporal drifts with the help of the grid density clustering approach. The grid density clustering approach tracks the data by maintaining the set of the classifier models with each cluster. The Structured weighted aggregation predicts the sample for the clusters. For the development of better classification environment, the grid density clustering approach was used along with the uniform sampling approach. The uniform sampling allowed the selected samples to be labeled correctly for handling the spatiotemporal drift detection. This created a better classification environment through the maintenance of the labeling budget. The two real world data streams such as the MAGIC and the EM dataset were used for studying the performance of the proposed model. The performance measure showed that the proposed model outperformed another model with 25.15% higher WSF. The proposed SE-PLS methodology used along with the uniform grid sampling scheme and the grid representation has provided a good subset of the labeled samples. The data stream with the label count of 10 was used for the classification. This technique, when used for the unbalanced data stream, will produce low results.

Neural network based algorithms

Ghazikhani et al. have proposed an online Neural Network (NN) model. The proposed NN model has two functions [4]. They are forgetting function and the specific error function. These functions were used for handling the concept drift and the class imbalance respectively. These functions provide a various rating for the error in separate classes to reduce the imbalance. The concept drift and the class imbalance occur in the data stream in the dynamically varying environment. The proposed model performs the data stream classification of the three synthetics and eight real-world datasets. The forgetting function handled the concept drift by combining with the proposed NN model. The error back-propagation algorithm allowed the detection of the concept drift. The proposed NN model produced an error function which contained the information about handling concept drift and class imbalance. The proposed model used the geometric mean metric approach when the imbalanced datasets are needed to be classified. The NN model classified the data stream both in overall and the incremental mode. The error function of the NN model has two parameters namely; m and n. Experiments were performed with the standard datasets to obtain the results of the parameters m and n. The parameter m and n show the imbalance ratio in the data stream. The performance result shows that the proposed method improves when the imbalance ratio becomes higher. The NN model proposed does not effectively identify the concept drift in the data stream.

Andreu et al. have proposed a thesis called supervised neural constructivist system (SNCS) for the data stream classification [7]. The proposed system subjects for the classification of the data in both static and dynamic environments. The proposed SNCS model is an accuracy-based neural-constructivist learning classifier system. This system finds the concept changes in the data stream by learning the data stream with the use of multilayer perceptions. The multilayer perceptions have a fast reaction capacity to identify the drift changes

in the environment. The proposed model learns the concept drift in the real time environment. They have proposed a supervised team miner for identifying the concept drift. The supervised team miner has a fast reaction capacity and a better adaptability to the dynamic environments. The Michigan-style LCSs technique is used along with the proposed model when the environment has harsh changes. The Michigan-style LCSs model learns the various changes in the dynamic environments. The proposed SNCS model when subjected to the varying standard datasets with real-world classification problems gives performance measure results. The proposed SNCS model has larger robustness against the complexities in the data mining, and it also has a remarkable reaction capacity to drift concept changes and noisy inputs. The SNCS has a better performance even when the padding variables in high dimensional spaces are employed.

Nicoletti have proposed an ensemble of CoNNs algorithm with functional expansion techniques for the data stream classification [8]. The functional expansion provides a better classification of the data stream with the concept drift. The performance results show that the when CoNN algorithm is operated in a non-stationary environment, it has better accuracy and a low computational cost. For a dynamic ensemble, the CoNN algorithm acts as a base classifier since it is independent of a model selection phase. The CoNN is a fast trained algorithm. For the data classification in the non-stationary environments, the FLANN algorithm has a good performance. Hence, The CoNN algorithm along with the FLANN algorithm performs the data classification in the dynamic environment. The CoNN algorithm as a base classifier along with the functional. Expansion when used for classification of the standard data sets such as SEA ensemble it shows better performance results. The advantage of the CoNN technique is low computational cost and improved efficiency. This algorithm when subjected for multiclass classification and regression then it has reduced performance.

Fuzzy based approaches

Shahparast et al. have proposed a scheme that assigns a weight to each fuzzy rule which can be used in the data classification [5]. An online method to adjust the weight of fuzzy classification is proposed in this thesis. The data set in the dynamic environment uses the fuzzy rule for the better classification scenario. The proposed rule-weighting classifier easily detects the drift and shift in the concept of data along the data stream. Their algorithm is the modified version of the existing batch mode rule-weight learning algorithm. The weight given to the fuzzy rule depends on the characteristics of the data streams. Implementation of this proposed algorithm in real life and some synthetic datasets have provided the performance measure statistics. The proposed classifier model is updated with the modified input data with the weights of fuzzy rules. Fuzzy rules contain the series of subsets. Hence, the weights of all fuzzy rules are updated by modifying the weight of all small subset present in the fuzzy rules. The proposed classifier model is updated based on weight modification in the fuzzy set. The weight adjustment of the subset in the fuzzy rule makes the proposed classifier update to be computationally efficient when dealing with varying data streams in the dynamic environment. The extended the eClass algorithm with added capability of rule-weight learning evaluates the performance of the proposed algorithm. The E-class algorithm uses online data streams of both the synthetic and real-world environments for the classification. Experimental results reveal that the proposed weight learning algorithm through the fuzzy rule improved the performance of the basic eClass algorithm for dynamic data stream classification. These algorithms can be compared with the other state-of-the-art methods for handling online data streams. The proposed online method does not detect the recurring concept drifts.

Isazadeh et al. have proposed in this thesis that online fuzzy data classification with the concept drift makes the classification to be difficult when the dynamic environment is used [6]. They have proposed the extend the Flexible Fuzzy Decision Tree (FlexDT) algorithm with multiple partitioning of the decision tree to provide the automatic classification. The proposed method aims to provide the balanced accuracy and tree size for the classification of the data in the data stream mining. In this proposed model, the updated model classifies the data stream by predicting its True and False detection. The class contains both the true and the false detection. In the real-time environment, classifier predicts the true class of each incoming instances. The performance parameters such as accuracy, tree depth, and the learning time are significant factors which influence the credibility of the proposed model. The proposed FlexDT can be extended to the MFlexDT which is more helpful in finding the continuous changes in the tree through adaptation of the incoming instances. The MFlexDT has an obvious advantage of having flexible classification when the environment is with the dynamic data. The temporary branches allow the controlling the noise and drift concept changes for the data stream in online systems. The proposed model provides an adaptive and flexible classification environment since the noises are eliminated. The proposed MFlexDT architecture is a practical architecture for applications such as wireless sensor networks. The proposed model requires more time for data classification.

Shaker et al. have developed an evolved version of fuzzy pattern tree learning for the data stream classification [9]. The current model adapts to changes through statistical hypothesis testing method. The fuzzy pattern allows the classification in the dynamic environment. The proposed model learns from the data stream and forms the extended model. The Fuzzy pattern used batch learning algorithm for pattern tree. This developed algorithm adapts for the problem of binary classification. The proposed eFPT meets the requirements of the incremental learning process in the data stream classification. The eFPT maintains a set of neighbor trees for replacing the current model. The neighbor trees allow the replacement by learning the characteristics of the data stream. The alternative tree realizes the current model through the replacement. The sliding window of fixed length continuously monitors the replacement decision made on the tree. The proposed algorithm has better accuracy in data classification with efficient identification of the concept drift. The production of the larger models of the eFPT can be an advantage when the target concept to be learned complex, and the resulting data classification process is sufficiently stable. The proposed eFPT when used for multi-class classification produced reduced efficiency.

Baruah et al. has proposed a fuzzy rule-based (FRB) classifier for the online data stream classification [10]. The proposed algorithm has an inherently adaptive nature as it learns the incoming example and changes adaptively. The proposed FRB classifier depends on the input and output of the classification model. The proposed FRB classifier obtains its antecedent parameters and rules through the partitioning of the input or input-output data. The classifier model distinguishes the data shift from the noise and appropriately handles noise in the online data stream. The proposed model classifies the data more efficiently in the online streaming of the data. The model gets updated in the real-time. The proposed FRB model gets updated to the DEFC model. The DEFC model adapts to the dynamic changes in the online data stream and is more robust to noise in the environment. The proposed

model performance evaluation is done by implementing the model in the standard datasets in the dynamic environment. In the online data streaming, the bandwidth parameter suffers the most variation. The adaptation of the parameter bandwidth in the model increases the efficiency.

Naive Bayes based algorithms

Gomes et al. have proposed mining recurring concepts in a dynamic feature space (MReC-DFS) in this thesis. One of the major challenges in the data stream classification is the learning recurring concepts in a dynamic feature space [11]. The multiple scanning of the data stream increases the historic data, thus an increase in the memory requirement. The proposed model detects and adapts to the model by learning the process and the contextual information in the data stream. The proposed model makes use of the desired percentile of the scores generated using the feature evaluation method with a various threshold. The proposed model adaptively selects the most relevant features in the information table and the threshold features. The proposed model implements the classifier model by running a test environment with the known dataset stream. The sued dataset contains the recurring concepts drifts from text mining. The parameters such as accuracy and the resource utilization evaluate the performance of the proposed model. The feature selection process in the proposed model minimizes the cost associated with learning recurring concepts in data streams in the dynamically varying environment. The proposed model achieves the tradeoff between accuracy and resources in the model and thus detects the concept drift efficiently. The proposed MReC-DFS model has less resource utilization in the dynamic environment. The proposed model has less performance in the online data stream classification.

Karabatak has proposed the Weighted Naive Bayes classifier (weighted NB) for the application on breast cancer detection in this thesis [12]. The proposed model depends on the Naive Bayes classifier theorem since it is one of the simplest classifiers. The breast cancer database provided the required data stream for the classification. The performance of the proposed model depends on the parameters such as sensitivity, specificity, and the accuracy values. These parameters find out the true and false detection of the data stream classification. The proposed model implemented in the cancer database provides the various performance results. The parameters sensitivity, specificity, and the accuracy values have obtained a value of 99.11%, 98.25%, and 98.54% respectively for various conditions of the data stream classification. Comparison of the results with the existing model proves the efficiency of the proposed Weighted Naive Bayes classifier model. One of the advantages of the Naive Bayes model is that they have a simple structure. Hence, the weighted NB classifier can be easily updated. The proposed model uses the grid search mechanism to find the optimum weight values for the weight assignment in the classifier model. The grid search mechanism used was computationally expensive, and they are application dependent. The initialization of the weight of the model is a complex task. This makes the algorithm to be inefficient when used in the dynamic environment.

Hoeffding-ID based algorithms

Yin et al. have proposed an improved Hoeffding tree data stream classification algorithm for the data stream classification in the dynamic environment [13]. The proposed model depends on the Hoeffding-ID algorithm. This algorithm effectively detects the intrusion detection field in the data stream. The proposed model divides the data stream to the data samples to provide better classification scenario. The number of data samples in the data stream depends on the limiting factor. In the extreme case, the proposed model imports the limiting factor to manage the number of the data samples. The intrusion detection field detects the occurrence of concept drift in the data stream. The proposed model finds the memory usage rate, True and False detection percentage of the data stream classification using the Hoeffding-ID. These results when compared with the existing models, the proposed model provided better results. The proposed Hoeffding-ID model has the good performance of TP rate and time of constructing the decision tree for the classification is very low when compared with the original Hoeffding model. The advantages of the proposed model are high TP rate, a low FP rate, and extremely low memory usage rate. The proposed model uses more feature for the data stream classification and thus has created a negative impact on the performance of the detection process. The proposed Hoeffding-ID algorithm has a very high FP rate for the data stream with the concept drift.

Sun et al. have proposed a classification model based on random decision tree in this thesis [14]. The proposed model determines the split point in the data stream with the use of Hoeffding Bounds inequality and information entropy instead of the random selection method. The Hoeffding Bounds determines the threshold inequality for the proposed model to detect concept drift in the data stream. The proposed model concentrates on the Coal Mine Safety Evaluation. The proposed random decision tree depends on the Hoeffding ID. The proposed model detects the concept drifts by evaluating a fewer number of the data samples. The performance of the proposed model is evaluated with factors accuracy. The proposed model detects most of the concept drifts in the data stream than the existing models with the consumption of fewer data samples. The proposed model, when implemented in the standard data sets, provides better classification scenario. It also provides a good classification of the simulated data sets. The higher accuracy of the data classification determines the safety evaluation of coal mine. The coal mine data stream has missing data and abnormal data contents, which makes the classification process to be difficult. The disadvantages of the proposed model are an inefficient classification of the incomplete label and noise problems in coal mine data stream.

KNN based algorithms

Perez et al. have presented the online classification algorithm for the dynamic data stream [15]. The proposed algorithm learns a Mahalanobis metric in an online data stream using similarity and dissimilarity constraints of the data. The proposed algorithm combines two approaches. They are Mahalanobis distance metric learning algorithm and a k-NN data stream classification algorithm with concept drift detection. The Mahalanobis distance metric learning algorithm allows the model to learn the data stream in the online and the k-NN data stream classification algorithm performs the required classification process based on the learning model. The streaming evaluation methodologies evaluate the performance of the proposed model by comparing it with the existing model. The proposed model has better concept drift detection than the other models. The parameter scalability ensures the incremental learning process in the data streaming to be effective. The learning process forms a Mahalanobis matrix for studying the characteristics of the data stream. The KISSME, an online optimization algorithm computes the similarity and dissimilarity matrices present in the Mahalanobis matrix from the vector product defined. The eigen decomposition value of the Mahalanobis matrix is the difference in the inverse of the similarity and dissimilarity matrices. The factors such as prequential error, prequential accuracy, and Q with

fading factors determined the performance evaluation measure of the proposed model.

De'Souza et al. has proposed a classification model for the Insect Recognition [16]. The proposed model evaluates the data stream with the concept drift. The classifier model gets regularly updated with the insect recognition in a data stream. The insect recognition data stream is large. The proposed model classifies the data stream by taking small data samples. The training set makes the data classification to be easy. Insect Recognition data stream is a real-time data set. From the initial set of results, the philosophy of inserting and removing examples from the training set are found to be effective in data stream classification. The adapted training set has a prime importance in the data stream classification. Inserting examples with the high classification confidence increases the accuracy of the data stream classification. The proposed classification model adapts to the change in the data drift. The climate change causes the concept drift to be occurred more gradual than in real conditions. The laboratory conditions are induced the concept drift in the data stream due to minor temperature changes, insect circadian rhythm, and aging. Various philosophies such as discarding training examples in the data stream for instance: by age, by some training samples of each class evaluates the performance of the proposed model. Artificial datasets make stream learning algorithm to be ineffective as they don't know the time of exact occurrence of the concept drift.

Probabilistic-based algorithms

Oliveira and Batista have proposed an incremental algorithm for data stream learning process [17]. The proposed method Gaussian Mixture Classification Algorithm (IGMM) is based on Gaussian Mixture Model. The proposed model finds the classification of the algorithm by defining the probability class for the data stream. The Gaussian Mixture Model makes the classification to be efficient in the dynamic data stream. The performance evaluation parameters measure the compatibility and the accuracy of the proposed classifier model. The proposed model shows better performance when compared with the existing models with the Gaussian approximation. The proposed model removes the outdated components with the explicit policy and constantly updates the model for the data stream with the concept drift. The proposed algorithm controls the maximum number of components in the data stream and has improved efficiency than the other models. The further improvement in the accuracy value is done through replacement of full covariance matrix with a diagonal covariance matrix in the classifier model. The preset parameter T used in the algorithm has a simple construction. The performance enhancement through the removal of the component by age and use of classification makes classifier to be more efficient. Some sophisticated approaches such as component fusion allow joining two or smaller components into one. This reduces the complexity of the component removal. The Gaussian component increases the complexity of the data stream classification.

Bargi et al. have proposed a novel adaptive online system for the data stream classification and segmentation [7]. The proposed model has the properties of the Markov switching models with hierarchical Dirichlet process priors. The proposed model does the classification with the probabilistic approach. The proposed adaptive online approach segments and classifies the data streaming over infinite classes. The model meets the memory and delay constraints of online streaming contexts. The model is further improved by including the predictive batching mechanism in the classifier. The predictive batching mechanism subdivides data streams into batches of variable size, through the imitation of the ground-truth segments. The proposed model performs the classification of the two video datasets in the real time. The experimental results show that the model is significant in frame-level accuracy, segmentation recall, and precision parameters. The classifier model determines the accurate number of classes in satisfactory computational time. They have extended the work by increasing the infinite class sets for the data stream to meet the finite memory requirements of the data stream. The predictive batching plug-in enhanced the proposed model by evaluating the variants. The proposed model with the efficient boundary prediction mechanism estimates the segment boundaries of the data samples by terminating the batch of data sample at the boundary points. The proposed model has more degrees of freedom and observes the data stream on common datasets.

Decision tree based approaches

Rutkowski et al. have proposed a method based on the Gaussian Approximation for the data stream classification [8]. The proposed algorithm has a solid mathematical basis and hence it is computationally efficient. The proposed model outperformed the existing Mc Diarmid Tree algorithm. The proposed model constructs a decision tree with a finite data sample of high probability from the data stream by choosing the best attribute. This is the same as the decision tree constructed with the entire data stream. The constructed tree is called the Gaussian Decision Tree (GDT). The constructed decision tree discusses various data stream mining problems. The node splitting allows the classification of the data. The proposed method performs the data classification based on the Taylor's Theorem and the properties of the normal distribution.

Gajbhiye and Vaidya have proposed a Novel Class Detection of Feature for mining concept-drifting in the data streams [18]. The proposed model uses the weighted ensemble classifiers for the data classification. The weighted ensemble trained the classifier model based on the data stream. The proposed model deals with the classification of the Feature-Evolving Data Stream to make the information system to be secured. The data stream with the high dimensional data has more rigid classification problems. The state-of-the-art data mining techniques used for the classification of these data streams have high qualities. The proposed model reduces the false alarm rate and increase detection rate in the data stream. The proposed model deals with the large data stream classification. The flexible decision boundary defines the decision boundary for outlier detection by allowing a slack space. The model enables the classifier to detect more than one novel class in the data stream at a time. Thus multiple novel detections allowed the classification of the data stream to be faster. The novel class detector in the classifier model detects the concept-drift and concept evolution present in the large data stream. The proposed model is used along with the graph-based approach. The graph-based approach distinguishes multiple novel classes in the data stream and defines the attributes. In this thesis, a model is used for classifying the unlabeled data and detecting novel classes in the data stream. The proposed SVM method effectively retrieved the data from input data content in the data stream. The proposed model has advantages in the effectiveness and less timely retrieval of the document.

Mirzamomen and Kangavari have proposed a neural network based approach for the data stream classification [9]. The proposed model has decision trees with the internal nodes contain trainable split tests. The proposed approach contains a trainable function based on multiple attributes in the internal nodes. This function provides the flexibility needed in the data stream context to perform better classification. The trainable function improves the stability of the algorithm. The proposed

model uses the fuzzy set along with decision tree to perform the data stream classification. The proposed evolving fuzzy min–max decision tree (EFMMDT) learning algorithm contains a decision tree with an evolving fuzzy min–max neural network. When the data stream is large, the proposed EFMMDT algorithm divides the instance space non-linearly based on various attributes of the information table content. This results in the formation of smaller and shallower decision trees in the classifier model. The proposed EFMMDT learning algorithm has an incremental, hybrid and one-pass algorithm for learning decision tree structure with data streams in a dynamic environment. The proposed model has a good equilibrium between plasticity and stability of the data stream. The EFMMDT algorithm has good accuracy and kappa statistics when compared to the existing algorithms. The performance of the proposed model is evaluated by implementing the model on the standard data streams with the concept drift. The proposed model was extended for studying alternative trainable split tests in the decision tree.

Other approaches

Mena-Torres and Aguilar-Ruiz have introduced a model named Similarity-based Data Stream Classifier (SimC) for the data stream classification [10]. The proposed model performs the data classification through the novel insertion/removal policy. The proposed model maintains an envoy with a small set of examples and estimators for guaranteeing good classification in the data stream. The proposed SimC model detects the class labels in the updated model and removes the unnecessary classes during the running phase of the classification process. The removed class possess zero value to the classification process. The parameters such as efficacy and efficiency evaluate the performance of the proposed model. The parameter efficacy determines the classification rate of the process and the term efficiency determines the online response time required for the classification process. When the proposed model is implemented in the benchmark data sets, the performance evaluation results suggests that the proposed model has better accuracy and the online response time than the other existing models. The proposed model performs the classification in the data stream where an envoy and a small sized set of information is kept to conserve the distribution of classes. The SimC uses a knowledge model to manage the size of the data stream. This knowledge model makes the insertion/removal policy to be easy. SimC achieves a proper balance among the classes of the data stream and guarantees the effectiveness in the minority classes for unbalanced problems. The performance evaluation of the model is done using the Friedman's test. The test results show a comparative performance.

Xua and Wanga have proposed a machine learning algorithm for the data stream classification [11]. The proposed model uses the ELMs as the base classifier for the classification process. The classification process makes use of the adaptive changes in the number of the neurons in the hidden layer of the classifier model. The proposed model has activation functions obtained through the random selection from a series of functions. The algorithm finds the decision results for unlabeled data by the weighted voting strategy from the trained data of the classifiers. This thesis proposed a fast incremental extreme learning machine algorithm called the IDS-ELM for the data stream classification at the dynamic environment. The proposed model is based on the fast search method. When the data stream had the concept drifts, then this algorithm uses the learning model for the classification process. The proposed IDS-ELM model has an improved classification accuracy and speed than the existing models. The proposed algorithm has a good stability when the sliding window changing occurs.

De'Souza et al. have proposed the MClassification method to classify the data streams with infinitely delayed labels [19]. The proposed model classifies the data stream in the online. The proposed method used the Micro-Cluster representation from online clustering algorithms to classify the data streams. If the concept drift is present in the data stream, then the algorithm uses the distance-based strategy to maintain the Micro-Clusters in the model. The performance evaluation in several synthetic and real data sets shows that MClassification has better accuracy results than the existing models. The proposed model has less computational cost than the existing model. The proposed model does not need critical parameters for data classification. The proposed model deals with incremental drifts occurring in the data stream. Detecting the incremental drift is more challenging than detecting the abrupt or gradual drift, given the significant overlap between concepts for a short period. The proposed algorithm in this thesis shows competitive accuracy results to the existing state-of-the-art methods regarding the practical time costs and a single parameter. The MClassification method does not need the critical parameter prior knowledge. The proposed model implementation in the data stream with the outliers has reduced performance.

Lughofer et al. have proposed two ways to recover economy and applicability of drift detection in the data stream [12]. They are semi-supervised approach and fully unsupervised approach. The semi-supervised approach employs single-pass active learning filters to select the most appealing samples in the data stream. The fully unsupervised approach uses an overlap degree of classifier's output certainty distributions for the data classification. The proposed approach simplifies the data classification process with the use of a weighting factor. The proposed model uses the modified version of the Page-Hinkley test to detect consecutive drift occurrences in a data stream. The application of the proposed model in the real-time data stream gives the performance evaluation parameters. The proposed semi-supervised approach detects three real-occurring drifts in data streams when applied in the standard data set. The proposed model uses the machine learning approaches for detecting the concept drift. This makes the algorithm to be less complex. The algorithm implementation in the on-line visual inspection problems with active learning framework gives better accuracy than the other algorithms. The proposed semi-supervised approach uses an active learning scheme which selects the samples from time to time for labeling with the use of the AL buffer. In the complete unsupervised approach estimating classifier's performance such as accuracy purely depends on the input features of the data stream samples obtained in the online. The proposed model does not use the fully labeled instances for identifying the concept drifts in the data stream. The proposed algorithm has more stability and accuracy than the existing algorithms. Li and Yu have proposed an extension of the classical SVM technique for classification of the online streaming of the data [13]. The proposed on-line classier (OLSVM) when implemented for the data classification in the health monitoring provides better results than other algorithms. The SVM model classifies the big data stream into a smaller size for the simplification. In the on-line structural health monitoring, large data streams serve as a disadvantage. The proposed OLSVN uses a recursive method to calculate the kernel of classical SVM. The application of the proposed model in the real-time health monitoring data provides the performance evaluation parameters. The experimentation of the proposed model using the lab-scale prototype shows that it can detect the damages and irregularity in the data stream. They have constructed the method for large data classification, during the transformation of the data set to a data stream. The comparison of the proposed model

with the existing on-line classification model shows an improved value of the accuracy. This OLSVM overcomes the problems of classical SVM technique disadvantages such as sluggish training of the data, the vast dimension of the kernel, and low classification accuracy. The recursive method of calculating the kernel of the model makes the model compact and less complex. The support vector machines make the data learning process to be slow.

Wang et al. have proposed a method based on the online learning framework for classifying the large data stream [14]. The existing online algorithms based on the first order online learning have reduced accuracy. Hence, in the proposed model extends the first order online learning to the second order for improving the classification process. The proposed Sparse Online Classification (SOC) method allows the classification of the data stream to be done in the second order learning model. The proposed model used the first order online learning model as a special case to derive the second order online learning model. The use of the second order information from the first order learning model makes classification of the data stream to be more accurate. The parameters such as efficacy and efficiency measure the performance of the proposed algorithm. The proposed model shows a better performance than the existing models regarding the efficiency and efficacy. The proposed sparse online classification (SOC) for large scale data stream classification has better efficacy and accuracy. The proposed model with the use of diverse online learning algorithms detects the concept drift more efficiently.

Feng et al. have proposed a data stream classification algorithm based on the Self-Representative Selection process in this thesis [20]. The proposed Incremental Semi-Supervised Classification approach via Self-Representative Selection (IS3RS) performs the data streams classification by analysing both the labeled and unlabeled dynamic samples in the data stream. The proposed IS3RS model finds the better representative exemplars from the sequential data chunk in the data stream. The selected exemplars from the data chunk allow expansion of the training set. This is done through the incremental labeling of the exemplars. The experimental evaluation of the proposed model on the benchmark data sets shows the effectiveness of the proposed IS3RS model. The proposed method uses the representation learning theory to find a subset of data in the data stream that efficiently defines the entire data set. The test subset of the data from the data stream improves the classification process with the learning model. The Representative learning model extracts the informative exemplars from the data stream and co-training technique labels the exemplars. The proposed model has effective accuracy and efficacy than the other online models when implemented on the benchmark datasets since incremental learning is used. The results show that the proposed model finds the informative exemplars to increase the training set and gradually find new classes to detect the change in concept drift in the data stream. The proposed algorithm has applications involving large data streams such as stock forecasting and the data mining tasks. The proposed method was further extended to a data stream with distributed version and a realization on a parallel computing platform.

Wozniak et al. have proposed a classification model with the learning algorithm for the data stream classification with the presence of concept drift [21]. The proposed uses a novel classifier training algorithm which is based on the sliding windows approach. The sliding windows approach allows implementing of forgetting mechanism in the data stream classification. Whenever a concept drift is detected in the data, stream the classifier model proposed gets updated. The forgetting mechanism does not consider the old objects come from the outdated model in the updated classifier model. Only a few of the examples from the data sample gets labeled since the forgetting mechanism uses the limited budget for labeling the examples. The learning model in the proposed method selects the defined examples for the labeling process. This process is important for the effective classification of the data. The model uses a learning paradigm for choosing the examples in the data stream to be labeled. The proposed model uses both the semi-supervised and the fully supervised approach for the data classification. The active learning algorithm in the proposed data stream classifier has better performance results when implemented on the standard datasets. They have presented the classifier model based on the active learning to detect the exact location of the data drift occurrence. The proposed model implementation on the several benchmark data streams states that the model can adjust to changes in the data stream. The parameters accuracy and efficacy evaluates the performance. The semi-supervised approach based on active learning has returned the same results as the fully supervised approach in the online data stream. The proposed model suffers from the disadvantage when the data stream is large and the use of the ensemble algorithms in the learning process.

Nguyen et al. have proposed a model for detecting line events in a wide-area power grid [22]. The proposed model classifies the data stream through the machine learning techniques to PMU data. The Smart Grid placed over the data stream environment simplifies the classification process. The proposed model uses an archived synchrophasor data from PMU located in the Pacific Northwest to build a decision tree for the classification environment. The proposed model shows better classification performance than the existing J48 algorithm. The model uses the data from PMU with a large and active power grid. The proposed model considerably outperforms hand-coded rules when identifying line faults from a distance. In this thesis, the algorithm focused on repeatedly detecting and classifying line events in online and archived PMU data using machine learning by replacing hand-built classification rules is discussed. The domain experts have developed hand-built classification algorithm. The hand-built classification algorithm finds complexity in the implementation process. The learning algorithms for the event classification have support vector machines for the event classification. The unsupervised clustering techniques to PMU data reduced the performance of the proposed model. The supervised learning process aims in identifying the unknown events and signatures of the online data stream. The proposed model found the application in the data cleansing and spoofed signal detection processes.

Rosa et al. have introduced an algorithmic approach for nonparametric learning in data stream classification [23]. The simple local classifiers used in the proposed model allow the learning model to update and classifies the data stream effectively. The learning model adapts to the changes in the data stream. The proposed model has the balance between the model complexity and predictive accuracy of the data stream. The parameters such as accuracy and the model complexity evaluate the performance of the proposed ABACOC Algorithm. The data classification performance depends on the positioning of the balls in the data mining model. The examples placed in the data stream allow the relocation of the ball centers for the further improvement of the algorithm. If the process develops incrementally, then the model requires more balls than the normal requirement. The proposed model keeps the model size bounded even in the presence of an arbitrarily long stream. They have presented a set of algorithms for nonparametric classification of data streams in the dynamic environment. The proposed model allows nonparametric classification, incremental learning, and dynamic addition of new classes, small model size, and fast prediction

at the testing time, essentially no parameters to tune in the data stream classification process. The proposed model implementation in the standard datasets shows the improved performance of the performance metrics. The proposed model has disadvantages in local dimensions of data that allow performing dimensionality reduction locally and incrementally.

Summary

From the above discussions, various data stream classification algorithms for classifying the data stream with concept drift was analysed. The discussion shows that the methods have faced more difficulties and challenges for classifying the data with the concept drift. The ensemble based models have suffered a disadvantage when used for spatial and temporal analysis of the data stream classification. Also, it does not consider the recurring concept drift in the data stream. This algorithm suffered a disadvantage during the use of the partially labeled streams. The fuzzy based approach disadvantage was the ability of reasoning and learning the classification model for various resources was less. This approach suffers a disadvantage with the use of the parallel streaming of online data. This approach, when used for the unbalanced data stream, will produce low results. The NN model proposed does not effectively identify the concept drift in the data stream. This NN model when subjected to multiclass classification and regression then it has reduced performance. The NN model requires more time for data classification. The NN model when used for multi-class classification produced reduced efficiency. The Naive Bayes model has less performance in the online data stream classification. The initialization of the weight of the model is a complex task. The disadvantages of the Naive Bayes model are an inefficient classification of the incomplete label and noise problems for online and large data stream. Artificial datasets make stream learning algorithm to be ineffective as the Naive Bayes does not identify the time of exact occurrence of the concept drift. The Gaussian component increases the complexity of the data stream classification. The Mclassification model implementation in the data stream with the outliers has reduced performance. The support vector machines make the data learning process to be slow. The Mclassification model realization on a parallel computing platform has less performance in the online data stream. The ensemble model suffers from the disadvantage when the data stream is large and the use of the ensemble algorithms in the learning process. The ABACOC model has disadvantages in local dimensions of data that allow performing dimensionality reduction locally and incrementally. Hence, to overcome the computational problems and perform the data stream classification more effectively Dynamic concept drift is best one.

References

1. Zhang P, Zhou C, Wang P, Gao BJ, Zhu X, et al. (2015) E-Tree: An efficient Indexing structure for ensemble models on data streams. IEEE Transactions on Knowledge and Data Engineering 27: 461-474.
2. Brzezinski D, Stefanowski J (2014) Reacting to different types of concept drift: the accuracy updated ensemble algorithm. IEEE Transactions on Neural Networks and Learning Systems 25: 81-94.
3. Gama J, Kosina P (2014) Recurrent concepts in data streams classification. Springer 40: 489-507.
4. Raja AMA, Swamynathan S (2016) Ensemble learning for network data stream classification using similarity and Online genetic algorithm classifiers. International Conference on Advances in Computing, Communications, and Informatics (ICACCI), Jaipur, India.
5. Canzian L, Zhang Y, Schaar M (2015) Ensemble of distributed learners for online classification of dynamic data streams. IEEE Transactions on Signal and Information Processing Over Networks 1: 180-194.
6. Sethi TS, Kantardzic M, Arabmakki E, Hu H (2014) An ensemble classification approach for handling spatio-temporal drifts in partially labeled data streams. 15th International Conference on Information Reuse and Integration, San Francisco, California, USA.
7. Andreu SA, Albert OP, Golobardes E (2014) Robust on-line neural learning classifier system for data stream classification tasks. E-Soft Computing 18: 1441-1461.
8. Junior B, Nicoletti MC (2016) Functionally expanded streaming data as input to classification processes using ensembles of constructive neural networks. 2016 International Joint Conference on Neural Networks.
9. Shaker A, Senge R, Hüllermeier E (2013) Evolving fuzzy pattern trees for binary classification on data streams. Online Fuzzy Machine Learning and Data Mining 220: 34-45.
10. Baruah RD, Angelov P, Baruah D (2014) Dynamically Evolving Fuzzy Classifier for Real-time classification of data streams. IEEE International Conference on Fuzzy Systems, Beijing, China.
11. Gomes JB, Gaber MM, Sousa P, Menasalvas E (2014) Mining recurring concepts in a dynamic feature space. IEEE Transactions on Neural Networks and Learning Systems 25: 95-110.
12. Murat K (2015) A new classifier for breast cancer detection based on Naive Bayesian. Measurement 72: 32-36.
13. Yin C, Feng L, Ma L (2016) An improved Hoeffding-ID data stream classification algorithm. The Journal of Supercomputing 72: 2670-268.
14. Sun G, Zhongxin W, Zhao J, Wang H, Zhou H, et al. (2016) A coal mine safety evaluation method based on concept-drifting data stream classification. 12th International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery, Changsha, China.
15. Perez JLR, Ribeiro B, Perez CM (2016) Mahalanobis distance metric learning algorithm for instance-based data stream Classification. International Joint Conference on Neural Networks.
16. D'Souza VMA, Silva DF, Gustavo AB (2013) Classification of data streams applied to insect recognition: initial results. Brazilian Conference on Intelligent Systems, Fortaleza, Brazil.
17. Oliveira LS, Batista G (2015) IGMM-CD: A Gaussian Mixture Classification Algorithm for Data Streams with Concept Drifts. Brazilian Conference on Intelligent Systems, Natal, Brazil.
18. Gajbhiye PR, Vaidya SG (2016) Classification and Adaptive Novel Class Detection of Feature - Evolving Data Streams. International Journal of Engineering Research and General Science 4: 616-622.
19. D'Souza VMA, Silva DF, Batista G, Gama J (2015) Classification of evolving data streams with infinitely delayed labels. 14th International Conference on Machine Learning and Applications, Miami, FL, USA.
20. Feng Z, Wang M, Yang S, Jiao-Key L (2016) Incremental Semi-Supervised Classification of data streams via self-representative selection. Applied Soft Computing 47: 389-394.
21. Wozniak M, Ksieniewicz P, Cyganek B, Kasprzak A, Walkowiak K (2016) Active learning classification of drifted streaming data. International Conference on Computational Science 80: 1724-1733.
22. Nguyen D, Barella R, Wallace SA, Zhao X, Liang X (2015) Smart grid line event classification using supervised learning over pmu data streams. 6th International Green Computing Conference and Sustainable Computing Conference, Las Vegas, NV, USA.
23. De Rosa R, Orabona F, Cesa-Bianchi NO (2015) The ABACOC Algorithm: a novel approach for nonparametric classification of data streams. International Conference on Data Mining Atlantic City, NJ, USA.