International Journal of Electrical and Computer Engineering (IJECE)
Vol. 8, No. 1, February 2018, pp. 19~25
ISSN: 2088-8708, DOI: 10.11591/ijece.v8i1.pp19-25

Concept Drift Identification using Classifier Ensemble Approach

Leena Deshpande, M. Narsing Rao
Departement Computer Science and Engineering, KL University, Vijaywada, India

Article Info

Article history:

Received Mar 9, 2017
Revised Jun 17, 2017
Accepted Dec 11, 2017

Keyword:

Accuracy
Classification
Drift
Ensemble
Frequent Pattern

ABSTRACT

Abstract:-In Internetworking system, the huge amount of data is scattered, generated and processed over the network. The data mining techniques are used to discover the unknown pattern from the underlying data. A traditional classification model is used to classify the data based on past labelled data. However in many current applications, data is increasing in size with fluctuating patterns. Due to this new feature may arrive in the data. It is present in many applications like sensornetwork, banking and telecommunication systems, financial domain, Electricity usage and prices based on its demand and supplyetc .Thus change in data distribution reduces the accuracy of classifying the data. It may discover some patterns as frequent while other patterns tend to disappear and wrongly classify. To mine such data distribution, traditionalclassification techniques may not be suitable as the distribution generating the items can change over time so data from the past may become irrelevant or even false for the current prediction. For handlingsuch varying pattern of data, concept drift mining approach is used to improve the accuracy of classification techniques. In this paper we have proposed ensemble approach for improving the accuracy of classifier. The ensemble classifier is applied on 3 different data sets. We investigated different features for the different chunk of data which is further given to ensemble classifier. We observed the proposed approach improves the accuracy of classifier for different chunks of data.

Corresponding Author:

Leena A Deshpande,
Dept of Computer Sci. &Engg,
K L University,
Green Fields, Vaddeswaram, Guntur District, A.P., India.
Email: deshpande.leena27@gmail.com

1. INTRODUCTION

Research over last few decades have developed many data mining algorithms for discovering knowledge underlying the data [1], [2]. These algorithms, however, are often used for static datasets, while recently developed new applications face the problem of processing large volumes of data generated in the form of data streams. Applications like sensor networks, web log analysis, and telecommunication systems require to process data generated at very high rates. Data stream imposes challenges like limited memory, less processing time and one scan of instances while processing. Traditional data mining algorithms cannot efficiently handle these problems, thereby leading to the development of stream data mining techniques. One of the challenges while learning from data streams is handling concept drifts, i.e., changes in data streams which deteriorate the accuracy of classifiers. This happens since classifiers learnt on past data instances are used for labelling recent data instances that reflect current concept which may be different from the old ones. Thus, for handling drifts in data streams, classifiers must use some technique for adjusting with changing environment [3-6]. Also, classifiers must be able to detect different types of drifts, sudden and gradual drift, that are characterized by the rate of changes observed [7].

Data stream is the sequence of data instances $\{x^t,y^t\}$ for time t=1,2,3,...,T, where x is set of attributes and y is class label. We assume that as data instance x^t arrives, classifier C predicts its class label. After some time, actual class label y^t is available and is used by classifier for evaluation and as additional information for training purpose. This technique called supervised learning is used by most of data mining algorithms. However constraints applied by stream data are not well addressed by this technique.As time elopes, the concept about which data is collected changes over time. This phenomenon, also called, concept drift is divided into two main categories as: sudden drift and gradual drift. The first type of drift occurs when source distribution S of data stream is suddenly replaced by another distribution S'. The later type of drift is associated with slower rate of changes in data streams. Typically, data instances from different source distributions start mixing, where probability of observing data instances from new source distribution increases and that of old distribution decreases over time Multiple algorithms have been proposed for dealing with concept drifts in data streams. Here we describe works related to our study briefly.

Drift detector is mechanism used for analyzing data instances and triggering alarm as soon as drift is observed. The trigger indicates need of rebuilding classifier. The most popular drift detection is Drift Detection Method in which predicted labels are compared with actual labels for determining classification errors. Classification error is monitored to check if it falls beyond certain threshold. When an error falls beyond threshold, alarm is signalled to store incoming data instances into a buffer. When alarm level is reached, new classifier is build on data instances in buffer and old classifier is removed.Concept drift is adapted into system when system is updated over current concept. The popular technique for accommodating current concept is windowing technique. This technique helps in keeping selected data instances in the system. Windowing technique is most widely used, since it keeps most recent data instances while eliminating data instances belonging to old concepts. Window size is common trade-off due to the fact that larger window size helps in keeping track of slower changes, but fail in case of sudden changes, whereas smaller window size can adapt sudden drifts efficiently as compared to gradual changes.The best way for dealing with concept drifts in data streams is ensemble technique which is a set of component classifier, votes of which are combined to predict class labels. Ensemble classifiers are best for dealing with changes in data streams due to their modular nature, which allows ensemble to be structured either by retraining component classifiers or by replacing weakest classifier by recently trained classifier or by updating weights assigned to component classifiers depending on their respective performances.

2. RESEARCH METHOD

Many algorithms have been developed with variations in basic processing technique of ensemble [8], [9] Minkuet. al. [10] proposed new approach that keeps different ensembles for dealing with diversity of concepts. It maintains different ensembles of data steams before concept drift and after concept drift in order to keep both old and new concepts in the system Senaratneet.al. [12] proposed a framework for determining hotspot of twitter activities and detecting drifts using kernel density estimation in streams of tweets. But it fails in determining the type of drift detected and taking measures over it. We overcome this problem by integrating online classifier and block-based classifier within single ensemble for reacting to different types of drifts efficiently.. To maintain minimum number of ensembles, we propose to use sliding window technique which helps in keeping most recent data instances. These data instances are used for retraining component classifiers, so as to keep ensemble updated over recent concept. In [13], [14] authors proposed a system that maintains ensemble of per-feature classifiers. One classifier is maintained for a single feature of a particular class. Such all per-feature classifiers of a class are combined for all classes making it hierarchy of weighted classifiers. The system spans over large memory space as the number of classes increases and thereby increasing time overhead. Our system analyzes features of class for checking out features responsible for drift, if any and for updating ensemble with necessary measures.

To overcome the problem of availability of actual class labels, learning technique is categorised into two types: online learning and block-based learning. In first approach, classifier predicts and evaluates as soon as datainstance is available. Whereas in later approach, blocks of data instances are used for evaluating classifier performance.Littlestone et. al. [15] put forward one of the algorithms for online learning, Weighted Majority Algorithm which aggregates predictions of component classifiers and updates weights of classifiers when predictions go wrong. Another ensemble proposed by Kotleret. al. [16] maintains set of classifiers, the weights of which are updates incrementally after each data instance. On each misclassification of data instance, the weights of the classifiers making false predictions are decremented.

Memory constraintis one among many challenges while handling data steams with concept drift. Hayat et. al. [17] proposed compact clustering technique to overcome this problem. Traditionally, every data instance of belonging to cluster was used for evaluating the cluster. The proposed compact clustering algorithm uses only neighbourhood instances for classification and clusters formed from unclassified

instances are compared with clusters of classified instances for checking abnormality and detecting drifts.Gao et. al. [18] proposed framework for detecting drift as well as new emerging classes in data streams using time constraints. The proposed system maintains buffer of instances unclassified by ensemble for certain time period. Instances remaining unclassified after time limit expiry are considered to be forming drift and are further analyzed for novel class evolution. Evolution of novel classes is another challenge in learning from data streams in which new classes emerge over time generating the need of restructuring ensemble by building new classifier for new class and eliminating the weakest component classifier. Novel class and feature evolution have been studied and many algorithms have been proposed for dealing with these issues [19-21].

The proposed system for handling concept drifts in data streams uses ensemble classifier technique for building base classifiers and using them for classification of testing data. The system builds online classifier as soon as new data instance is available and when block of fixed number of data instances is formed, block based c $w_i(t) = \left[\frac{1}{1 - A_i(t)}\right]^{\mu}$ classifier is developed. The classification of in coming data instance is done using weighted majority of base classifiers using weighting functions as:

Where, $w_i(t)$ is weight of base classifier C_i at time t and $A_i(t)$ is the accuracy of classifier C_i at time t.

The accuracy and error rates are monitored for each type of classifier continuously over blocks of data instances using Error Rate function as:

$$E_{ij} = (1 - f_{iy}(x))^2$$

Where, E_{ij} is the error rate of classifierC_i on recent block B_j of data instances and$f_{iy}(x)$ is the probability given by the classifierC_i that x is an instance of class y.

As the value of error rate monitoring crosses certain threshold, drift is detected. These drifts are analyzed and ensemble is updated accordingly.

3. RESULTS AND ANALYSIS

In our experiments, we evaluate our proposed ensemble that combines online classifier and block-based classifier. We implemented our ensemble system in Java. The experiments were performed on computer system with Intel Core i5 480M @2.67 GHz processor and 4.00 GB of RAM.

We tested the performance of ensemble with single component classifiers. Our ensemble used k=5 component classifiers; *NB Tree*, *J48*, *Logistic*, *Random Forest* and *Bagging*. The size of block used for all component classifiers and ensemble was equal *d*=100 as this size was best suitable for more accurate results. We evaluated ensemble performance for different sizes of block: 50, 100, 200, 500 and 1000. We observed that the statistical comparisons of performances of ensemble for each of above block size gives better results in terms of accuracy when block size was 100. However Block size does not alter ensemble accuracy significantly, but block size matters in case of drift detection. If the block size is large it ignores drifts that lasted for small time, while smaller block size detects drifts even if there are blips or noise in data streams.

Real world data contain no precise information about occurrence or type of drifts in it. So it is practically impossible to test the desire accuracy in terms of drift however a manual drift is to be inserted in the data to achieve the target.So we decided to use publically available machine learning benchmark datasets gathered in the UCI repository [22] that signified presence of gradual drifts.

We evaluate our ensemble of online classifier and block-based classifier against single classifiers as well/ We chose J48, NB Tree, Logistic, Random Forest and Bagging as component classifiers of basic ensemble. The ensemble is further modified for learning incrementally as well as in blocks of fixed size. From performance comparison of proposed ensemble with component classifiers as shown in Table 1, we can see that ensemble improves the accuracy of classification on all datasets and ensemble takes equal processing time.

Table 1. Performance Comparison between Component Classifier and Proposed Ensemble

Classifier	Classifier accuracy (in %) on Census income dataset	Spam email dataset	Electricity dataset
NB Tree	86.6957	83.4624	72.9652
J48	87.1901	85.0879	80.4749
Logistic	85.2062	84.4444	76.0979
Random Forest	84.4726	88.1553	81.9853
Bagging	87.1257	86.0137	83.3485
Proposed Ensemble	94.4094	92.6495	92.8426

Traditional approach of classifying testing dataset against given training dataset becomes inefficient while dealing with data streams, as the data streams keep arriving unboundedly and testing dataset is not completely available. The modification to ensemble by adding online and block-based learning component to classifiers shows significant role for keeping track of classification accuracies. Thus the two approaches: online classification and block-based classification are applied to solve this issue. These approaches help in monitoring classifier performances as data instances arrive and keeps track of changes in data stram.

In block-based technique, size of the block can be instance based or time based. In instance based technique, blocks of fixed number of instances are used. While in time based technique, blocks of data instances arriving over specific time period are used. As data streams arrive at any rate, it is not feasible to keep track of performance in time based technique, since some blocks will be densely populated while others being rarely populated. We observed this pattern in electricity data where demand and supply changes drastically with different time stamps We evaluated our ensemble using different sizes of block and compared our results with experiments conducted using MOA framework. The results are shown in Table 2. From the experiments we concluded, although performance of ensemble remains constant for any block size, the block size of 100 instances gives results for drift detection similar to those obtained using standard MOA (Massive Online Analysis) framework. Thus for further experiments we used 100 instances as block size.

Table 2. Performance comparison using different block size

Block size (in no. of instances)	Average accuracy of Ensemble	No. of drifts detected
50	86.0893	7
100	86.0810	3
200	86.0324	1
500	85.9928	1
1000	85.9574	0

Table 3 shows drifts detection using online and block-based method to ensemble. We analysed that for any dataset our ensemble works well in comparison with the component classifiers, online and block-based classification while detecting gradual and sudden drifts efficiently.

Table 3. Performance comparison using online and block-based technique

Classifier	Accuracy Comparisons using:			
	Online technique over:		Block-based technique over:	
	Census income dataset	Spam email dataset	Census income Dataset	Spam email dataset
NB Tree	81.4659	78.6537	85.5649	81.1856
J48	79.7652	72.7620	84.7610	78.6970
Logistic	78.3652	74.9539	83.9865	78.0652
Random Forest	83.6575	74.2886	83.4569	79.9099
Bagging	80.3315	76.2449	85.7326	83.6666
Proposed Ensemble	87.2546	88.0241	87.2701	86.1635
No. of drifts detected	-	4	3	1

Further we analysed the detected drifts to extract the hidden patterns in drifts that could potentially help us to find out the technique that can respond when next drift encounters. The analysis of drifts shows that values of few attributes are significantly missing and this leads to one of the root cause behind drift. The analysis of block before and after drift concluded that few blocks before drift, missing values of attributes started to observe in ascending way towards drift and in descending way afterwards the drift. The peak frequencies of missing values were noticed during drift. The results obtained as shown in Table 4 summarize that as the frequency of missing values started to increase, the performance of ensemble started degrading and visa versa. The values of attributes like workclass, occupation and native country were found missing during drifts.

The same pattern as shown in Table 4 was observed for all the further drifts in Census income dataset. This is recurring type of drift. Also we computed the information gain over the attributes of Census income dataset and we observed that the attributes that were found missing during drift are associated with highest values of information gain, thereby leading to be the main contributor to classification task. Thus handling missing values of attributes is also one of the key task in our proposed system.

We performed many experiments for handling missing values of attributes by either removing the instance that carries missing values or by replacing them using mean-mode imputation. We replaced the

missing values of numeric type of attributes by mean of values observed in previous instances and missing values of nominal attributes by mode. The graph shown in Figure 1 describes that replacing missing values by mean-mode imputation shows improved performance over other techniques: not dealing with missing values and removing missing values.

Table 4. Summary of detected drifts

Blocks observed	Count of missing values			Ensemble Accuracy over blocks
	workclass	occupation	native country	
10 blocks before drift	6	6	2	87.7510
5 blocks before drift	14	14	8	86.5461
Drift blocks	31	31	10	83.6993
5 blocks after drift	18	20	8	86.9491
10 blocks after drift	9	10	4	87.5421

Another problem that causes performance loss of classifier is the training dataset. Concept drift is the scenario where old concept occurs with less probability and new concept is mostly seen. In this situation, classifiers show degraded performances. This is because classifiers are trained on dataset that carries old concept and data instances that are being classified belong to new concept. Thusincorporating a new concept in learning process is an important task. This can be done in two ways: by using just classified recent block of data instances for retraining classifier ensemble or by adding recent block of data instances to training dataset iteratively and using it for retraining ensemble. We conducted different experiments in which for each case, we changed training dataset in three different approaches. In the first approach, we did not retrain ensemble once it is trained at starting with provided standard training dataset. In second approach, we first trained ensemble using standard training dataset and tested the first block of data instances. We used this recently tested block for retraining the ensemble and testing next block of data instances. The process is repeated iteratively for next blocks of data streams. This case of keeping recently classified block as training dataset maintains the current concept, but fails to keep track of old concepts as time proceeds. Thus in third approach, while proceeding forward we simply added recently classified block to the training dataset and then retrained the ensemble. Figure 2 shows the performance of ensemble using the above three approaches of retraining ensemble. The results concludes that adding recent block of data instances to provided standard training dataset is a wise solution and improves the performance of ensemble very well.

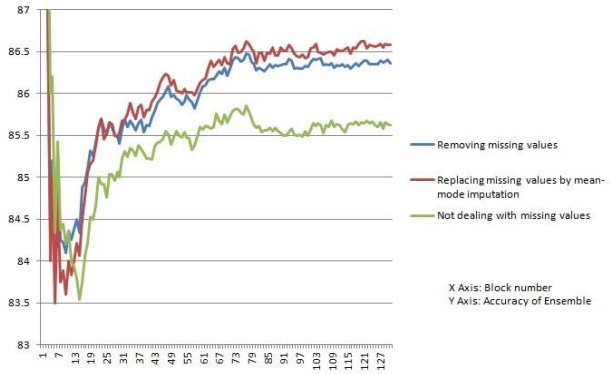

Figure 1. Performance comparison between different techniques of handling missing values

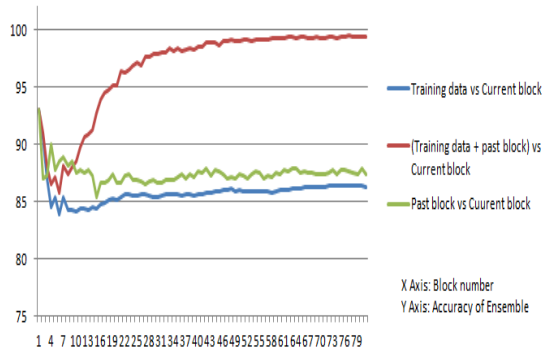

Figure 2. Performance comparison between different approaches of re-training ensemble

4. CONCLUSION

Most of recent applications generate large volumes of data at rapid rate, thus establishing the need of special data mining techniques for critical sensitive applications. Data stream mining is solution to the problem. Data stream mining poses challenges like limited memory storage, performance and change in concepts underlying data. Identification of two types of drifts, sudden and gradual drift may degrade the classifiers performance in terms of accuracy. To monitor such causes and parameters of drift identification, our proposed ensemble combines the characteristics of online classification technique and block-based classification technique and detects both types of drifts efficiently. Further our system analyzes theattributes

which are responsible behind the change in accuracy.Noisy, unbalance data and missing values of attributes were found to be the root cause behind the drift [23]. Our proposed system shows improved performance while detecting both kinds of drifts efficiently.However our work focuses on offline streaming data.

The contribution opens several directions for research studies. Current works in data stream classification focus detection of concept drift. Adaption of drifts to the system leads to new line of research. An interesting future work would be to identify evolution of new concepts for online streaming data. Recent techniques can be further extended for solving novelty detection problem.

REFERENCES

[1] J. Han, "Data Mining: Concepts and Techniques", Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2005.

[2] N. Littlestone, M.K. Warmuth, "The weighted majority algorithm", *Inf. Comput*. 108 (2) (1994) 212–261.

[3] M.M. Mausud, J. Gao, L. Khan, J. Han, and B. Thuraisingham, "Classification and novel class detection in concept-drifting data streams under time constraints", *IEEE Trans. On Knowledge and Data Engineering*, Vol. 23, No. 6, pp. 859-873, June 2011.

[4] Bifet, G. Holmes, Rr. Kirkby, B. Pfahringer, "MOA: Massive Online Analysis", J. Mach. Learn. Res. 11(2010) 1601-1604.

[5] J. Han, "Data Mining: Concepts and Techniques", Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2005.

[6] N.C. Oza, S.J. Russell, *"Experimental comparisons of online and batch versions of bagging and boosting"*, in: Proc. 7th ACM SIGKDD Int. Conf. Knowl. Disc. Data Min., ACM Press, New York, NY, USA, 2001.

[7] N. Street, Y. Kim, *"A streaming ensemble algorithm (SEA) for large-scale classification"*, in: Proc. 7th ACM SIGKDD Int. Conf. Knowl. Disc. Data Min., ACM Press, New York, NY, USA, 2001.

[8] N. Littlestone, M.K. Warmuth, "The weighted majority algorithm", *Inf. Comput*. 108 (2) (1994) 212–261.

[9] X. Zhu, P. Zhang, X. Lin, and Y. Shi, "Active learning from stream data using classifier ensemble", *IEEE Trans. On Systems, Man and Cybernetics- Part B: Cybernetics*, Vol. 40, No. 6, pp. 1607-1621, Dec. 2010.

[10] L.L. Minku and X. Yao, "DDD: A new ensemble approach for dealing with concept drift", *IEEE Trans. On Knowledge and Data Engineering*, Vol. 24, No. 4, pp. 619-633, April 2012.

[11] X. Zhu, P. Zhang, X. Lin, and Y. Shi, "Active learning from stream data using classifier ensemble", *IEEE Trans. On Systems, Man and Cybernetics- Part B: Cybernetics*, Vol. 40, No. 6, pp. 1607-1621, Dec. 2010.

[12] J. Han, "Data Mining: Concepts and Techniques", Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2005.

[13] Meenakshi Anurag Thalor, Shrishailapa Patil "Incremental Learning on Non-stationary Data Stream using Ensemble Approach", *International Journal of Electrical and Computer Engineering*, Vol 6, No 4: August 2016.

[14] B. Parker, A.M. Mustafa, and L. Khan, *"Novel class detection and feature via a tiered ensemble approach for stream mining"*, IEEE 24th International Conference on Tools with Artificial Intelligence, Vol. 16, No. 8, pp. 1171-1178, Nov. 2012.

[15] Maimon, L. Rokach (Eds.), "Data Mining and Knowledge Discovery Handbook", 2nd ed., Springer, 2010.

[16] L.I. Kuncheva, *"Classifier ensembles for changing environments"*, in: Proc. 5th MCS Int. Workshop on Mult. Class. Syst., LNCS, vol. 3077, Springer, 2004.

[17] M.Z. Hayat and M.R. Hashemi, *"DCT Based Approach for Detecting Novelty and Concept Drift in data streams"*, IEEE Conference on Soft Computing and Pattern Recognition, pp. 373-378, Dec. 2010.

[18] M.M. Mausud, J. Gao, L. Khan, J. Han, and B. Thuraisingham, "Classification and novel class detection in concept-drifting data streams under time constraints", *IEEE Trans. On Knowledge and Data Engineering*, Vol. 23, No. 6, pp. 859-873, June 2011.

[19] M.M. Masud, Q. Chen, L. Khan, C.C. Aggarwal, J. Gao, J. Han, A. Srivastava, and N.C. Oza, "Classification and Adaptive Novel Class Detection of Feature Evolving Data Streams", *IEEE Trans. On Knowledge and Data Engineering*, Vol. 25, No. 7, pp. 1484-1496, July 2013.

[20] M.M. Masud, J. Gao, L. Khan, J. Han and B. Thuraisingham, "Integrating Novel Class Detection with Classification of Concept Drifting Data Streams", ECML, Springer/PKDD, pp. 79-94, 2010.

[21] M.M. Masud, Q. Chen, L. Khan, C.C. Aggarwal, J. Gao, J. Han, and B. Thuraisingham, *"Addressing Concept-Evolution in Concept Drifting Data Streams"*, IEEE International Conference on data mining, pp. 929-934, 2010.

[22] Frank, A. Asuncion, UCI machine learning repository 2010.<http://archive.ics.uci.edu/ml/datasets>

[23] Fatma Karem, Mounir Dhibi, Arnaud Martin, Med Salim Bouhlel, "Credal Fusion of Classifications for Noisy and Uncertain Data", *International Journal of Electrical and Computer Engineering*, vol 7 No 2, April 2017.

BIOGRAPHIES OF AUTHORS

Leena A Deshpande received M E degree in Computer engineering in 2008. She is a Phd Scholar, working in data mining domain from KL UniversityVijaywada. Her current research interests include information retrieval, data mining, and Big data.

Dr.M.R.Narasinga Rao is a PhD in Computer Science and Systems Engineering from Andhra University, Visakhapatnam, India. He holds an M.Tech in Computer Science from Birla Institute of Technology, Mesra Ranchi, India. He is currently working as Professor in the department of CSE at KL University. He has published number of papers in national and international journals. His research interests include Applications of Neural Networks, Content Based Information Retrieval, Finite Automata and Bioinformatics. Guided number of students in graduate and postgraduate level courses and is currently having number of research scholars from KLUniversity. He is a recipient of Sastra Award for publishing papers in the year 2008 by Vignan Institute of Information Technology, Visakhapatnam.