AN EXPERIMENTAL EVALUATION OF CLUSTERING AND CLASSIFICATION OF HIGH-SPEED DIMENSIONAL DATA STREAM IN DYNAMIC FEATURE SELECTION

G. Senthil Velan, St. Peter's Institute of Higher Education and Research Chennai
K. Somasundaram, Chennai Institute of Technology Chennai
V.N Rajavarman, M. G. R. Educational and Research Institute University Chennai

ABSTRACT

There are many methods to alter the data flow theoretically and operationally. When new and extra features are introduced to the stream, or when the significance and relevance of a feature changes as the stream proceeds, a feature-level shift happens. This kind of shift has not garnered as much attention as conceptual alterations. Several clustering techniques (including density, drawing, and grid methods) utilise some kind of distance as a similarity metric, which is problematic with high-dimensional data, since the curse of dimensionality may lead to distance measurement, and any The ideas are extremely tough. calculate. We propose merging them and rephrasing them as feature selection issues, or more specifically dynamic feature selection problems, rather than attempting to answer each of these problems separately. We suggest utilising dynamic feature masks that vary over time to categorise big data streams. n Take action to group similar characteristics that have not been protected. When the perceived significance of characteristics changes, the mask will be changed to reflect the change; before that, smaller features will be unmasked, and comparable features will be disguised as required. In addition, the suggested technique is versatile and may be utilised with all popular density clustering methods, which lack a drift response mechanism and are impacted when dealing with huge volumes. Two texts and two picture sequences are utilised for assessment. The suggested dynamic function mask enhances the efficiency of grouping in all directions and lowers the processing time needed by the basic approach.

Keywords: A Clustering Technique for Data Streams, The Drifting of Features, Feature Evolution, and Unsupervised Selection

INTRODUCTION

Changes in addition to time and storage constraints are important issues when collecting data streams. It is important to detect and respond to changes for effective real-time analysis. Flux changes may vary. Let S [xt] be 0 stream. Where x is a two-dimensional t-vector. Let Y be a set of clusters identified by k: y1,. The point xi can be assigned to the cluster Y as a probability condition Pt (yj xi), a possibility of xi grouping Y into groups at time t. Conceptual development is a possible kind of change. This idea was created from the stream with the advent of a new new group, ym. This happens when the attributes of the data change. That is, in the basic process, it happens when x changes. These types of deflections are commonly referred to as Pt (x) virtual deflections. The second type of drift is actually called drift, or change in Pt (yx). For example, if the yj group is given a t point xi, then xi is assigned to the ym tβ group. For example, it can occur when yj and ym

Citation Information: Velan, S.G., Somasundaram, K., & Rajavarman, V.N. (2022). An experimental evaluation of clustering and classification of high-speed dimensional data stream in dynamic feature selection. *Journal of management Information and Decision Sciences, 25*(S2), 1-9.

clusters reference individual locations in functional space. The third unnoticeable fix is a functional level change. There are two ways to move and change attributes. Arrival instance x is in position dx f1,. , Assuming it is in Fd. The meaning of the characteristics is that the functional strength or connection is different for consistency reasons. For example, the importance of a particular word in text mining depends on the passage of time. For example, more words and characters will appear in a text string that changes dimension x as new features emerge. We have made a lot of efforts for operational integration, but when these adjustments occur, little investment has been made to improve performance. Also, many groups of technologies currently proposed (density oriented, graphical) are problem remote similar measurements for high-dimensional data because distance and density measurements are difficult to quantify. We propose to combine both problems and express in the problem of choice to solve both problems. Function selection (FS) tries to identify the subset of the most important characteristics F F (in this document, "function" and "dimension" are used interchangeably). F excludes duplicate functions (fif and fi/f) used for grouping of existing data. Because this can change over time, there may not be a highly confidential way of anomalous data. Significant changes will result in the deletion of existing clusters and the possibility of new clusters being detected in the current data. The two points are sometimes inseparable. There are many feature sets, for example when the number of features has changed 'significantly' (ft + 1) or the former is a required attribute (fi/ft but fi/fi/ft/1) has become insignificant. Because of these challenges, we recommend using dynamic performance masks to group big data streams. The transfer is run automatically FS after each window divided into β windows. Duplicate features are grouped by masking unmasked data features. The size of the mask changes as new items arrive one after the other. Grouped points include all features (not a subset of characteristics), but only a subset of related features in the grouping process. Therefore, we offer a new technology that uses dynamic masks to group large data streams. His main contributions are: Feature drift and development can be completely ignored and monitored, allowing you to evaluate features over time. ... this technique can be used for all current density combination methods that are algorithmic, have no major drift mechanism, and do not have a huge amount of data that is not commonly available. The proposed method can be applied to the present algorithm by reducing the time and increasing the accuracy. Part 2 gives an overview of the tasks involved, and Part 3 details the basic tasks. The fourth section shows the proposed approach. Section 5 summarizes the experimental results. Finally, we end the sixth segment.

LITERATURE REVIEW

Numerous research have been conducted on feature selection (FS), and the findings have been extensively discussed in (Tu, 2009; Bhatnagar, 2014). For the most part, research has concentrated on scientifically controlled ways of evaluating the significance of functions and categories. Choose the most distinguishing feature (or feature set) from among the available options. The FS technique may be broadly split into two types: filtration and packing. filtering Fisher estimate (Katakis, 2005), information retrieval (Roffo, 2015), and Pearson coefficient (Fahy, 2019) are some of the most used techniques. Boxing methods make use of models or sub-classifiers to examine subsets of functions, which is a kind of classification. GASVM (Li, 2018) is an example of a genetic algorithm that discovers subsets of characteristics and then utilises conventional support vector machines to assess the likelihood of these possible subsets being found. Packaging techniques and filtering methods are two types of uncontrolled procedures that may be distinguished. Uncontrolled packaging approaches make use of cluster methods to assess sub-feature sets (Rand, 1971), which are used to evaluate sub-feature sets. In most cases, this technique is computationally expensive, but it is compatible with the method developed by Alelyani, et al.,

Citation Information: Velan, S.G., Somasundaram, K., & Rajavarman, V.N. (2022). An experimental evaluation of clustering and classification of high-speed dimensional data stream in dynamic feature selection. *Journal of management Information and Decision Sciences, 25*(S2), 1-9.

This is referred to as the "chicken and egg problem" (Bhatnagar, 2014). Ideally, you should begin by searching for items before clumping them together. Alternatively, you should cluster things first before selecting them, attempting to group and pick objects at the same time. Examples include situations when data from the same category of data is routinely placed near the decision-making camera, in which case the inherent characteristics of the data are determined by uncontrolled filtering methods. On the basis of this assumption, the characteristics are chosen using either the local power or the Laplace exponent method. This technique was used in (Ester, 1996) and detailed in section 3 for non-recorded FS, and it is similar to that used in (Ester, 1996). Infinite selection of features (Roffo, 2015) selects features by taking use of the convergence properties of matrix power series. It is feasible to compare a subset of features using routes across various feature distributions using paths between different feature distributions. An algorithm for finding features that are capable of retaining their original data structure was suggested in (Fiscus, 1999), and it is described in more detail in (Forestiero, 2013). They utilise spectrum analysis methods to evaluate the connection between various functions, and then they apply their Multi-Cluster Selection Feature (MCFS) algorithm to pick the function that conserves the most energy. Section III provides a more in-depth description of this technique. The majority of FS research is conducted using static data packets, although more recent study has concentrated on data transmission FS. (Bifet, 2009) provides an overview of the most recent studies. FS. Katakis & Allabor (2005) are two of the most well-known figures in Greek history. It was among the first data streams to find a solution to the FS issue. The author shifts his attention to the issue of huge dynamic functional spaces in this section. You are utilising a text flow example that includes all of the words that are conceivable in the functional space to demonstrate your point. The number of texts that emerge increases the number of new words (features) that arise, and the range of known feature areas grows and changes. Statistical data accumulated over time In the Chiquare standard, the first n words of each text are chosen as the input for the NaiveBay categorization, which is based on the Chiquare standard. Update the statistics that have gathered. Top-level functions may be upgraded or degraded, and the classifier will make any necessary changes to the new function. The Heterogeneous Drift Set Feature (HEFT) (Nguyen, 2012) makes use of the Fast Correlatype filter (Hall, 1999) to choose the controlled filter technology with the best performance in each data flow window in the data flow window. The best features should be used to learn classifiers, and then they should be combined into a set, with each classification being trained on a distinct subset of features. Online weight ratings were utilised by o and Cohen (Carvalho, 2006) to determine the relative significance of each characteristic. It's worth noting that the researchers discovered that adding some of the lower levels may enhance the accuracy of the categorization. There are 90 percent of higher function and just 10 percent of lower function being utilised, according to the author. The notion of symmetric uncertainty is introduced in and is expanded to (based on the concept of information theory) for feature selection (discussion). A sliding window filter method, DISCUSS has nothing to do with classifiers and operates in the same way as a filter technique. They are chosen in accordance with a performance-based method, in which the perceived value of a feature subset is determined by the anticipated nature of the class and the degree of redundancy in the feature subset itself. It has been shown that using this selection technique may enhance the performance of two distinct courses. As well as others. Created adaptive gain FS (ABFS), which selects a function by combining gain (Freund, 1990) with a decision tree stump (a decision tree with a root node connected with a terminal) and a decision tree stump (a decision tree with a root node associated with a terminal). Reinforcement increases the importance of training samples that are difficult to categorise, while sample selection stumps are more difficult to classify than training samples. It has been shown by ABFS that categorization indicators have been improved while management expenses have been lowered. Other monitoring techniques, such as (Bifet, 2009; Gomes, 2017), make implicit selections

Citation Information: Velan, S.G., Somasundaram, K., & Rajavarman, V.N. (2022). An experimental evaluation of clustering and classification of high-speed dimensional data stream in dynamic feature selection. *Journal of management Information and Decision Sciences, 25*(S2), 1-9.

of characteristics. DXMiner (Masud, 2011) is a dynamic FS classification technique that is used in the mining industry. The method may make use of supervised or automated filters, depending on the situation.

METHODOLOGY

The suggested approach calls for an unattended function selection tor. We examine three current dynamic function mask static techniques. Each technique and the clustering algorithms that we employ to assess the suggested dynamic feature mask are explained below.

Unsupervised Feature Selection

Variance

The simplest but most successful technique for uncontrolled feature selection is the greatest variance method; the average square departure from the average feature value. X={x1,. ., xN } is a case of N, where xi=R d: Var(Xi)= 1 N X N j=1 (Xij − µi) Var(Xi) 2. 2. (1) A wider variance indicates that the characteristic has more representational power. The idea here is that it has limited predictive value when a characteristic does not change significantly (if it is almost always relevant for each class). However, if one characteristic for each class is sufficiently distinct, it may be better to discriminate across classes. The variance for each function is computed, the characteristics are classified in the order decreasing and top n characteristics are chosen. 2) LAPACIAN SCORE 2) In order to maintain the geometric form of the region under data protection, the Rapatian Rating was created for this purpose. A representation of this local structure is provided by the graph below, which is modelled as the graph below. • There are N nodes in the graph G that is next to this one. If the nodes xi and xj are nearby, an edge is drawn connecting the nodes I and j. • To represent the local structure, make use of the weighted G matrix S. Make use of the RBF function to weight the edges between nodes I and become a constant ttoR using the weighted edges between nodes I. 1 t=(ek (xixj) 2 1 Sij=(ek (xixj) 2 1 t: If I and j are near to zero (2) • G and the weight matrix S should be taken into consideration while evaluating the features of the system Importance. Make an estimate of the value of the response L by reducing Lr=P ij [fr] as much as possible. 2) 2Sij Var (fr) Lr=P frj) 3) Sij is good and big (so fri frj is small). Consequently, predatory assessment is a useful feature that should be used with care. The most up-to-date features have been chosen. 3) Select Multi-Cluster Attributes from the drop-down menu (MCFS) MCFS (Cai, 2010) makes use of spectral clustering to identify which structural characteristics are the most effective for protection. The top Lapas lattice vector is used to conduct the spectral clustering operation. Lapaci estimation, for example, involves drawing the closest graph G and weight matrix S. ... The S-data collector "stretches" smoothly to the "level" of data points and chooses features via "flat" integration, resulting in a more natural user experience. S is used to create the diagonal matrix D with the calorific value S. The diagonal matrix D with the calorific value S is formed as S. Dij is the same as Pj Sij. L=DW is derived by solving generalised problems, such as the "planar" integral of the data generated by these matrices, and then graphing the results. The following are some hypotheses on the characteristics of Ly=Dy Ly (4). This gives the result vector Y=Y1...Yk, where k is the number of data groups in the data set. If static location data is provided, the value of K may be modified to identify the most well-known pre-existing answer. In the case of transmission, k is unknown. Take a look at the previous window and make use of the number of groupings. In each feature, the MCFS scores are sorted in decreasing order of importance, and the top n are chosen.

B. STREAM CLUSTERING ALGORITHMS To test the proposed method, we use four density stream clustering methods: MDSC (Fahy) ACSC (Fahy, 2019) CEDAS (Hyde, 2017) and DenStream (Cao, 2006). MDSC (Fahy) ACSC (Fahy, 2019) CEDAS (Hyde, 2017) and DenStream (Cao, 2006). The term "cluster" refers to areas of high density separated by areas of low density. A microcluster is formed from a set of points that correspond to each other's proximity (Euclidean distance) in functional space. N points XEj, j=1 and N=2 define a microcluster with 4 components. These are: In the formula, N represents the number of microcluster points. Each point is two-dimensional. LS represents the point linear sum (*i.e.,* PN j=1 XE 2 j). SS represents the sum of the compressed points (*i.e.,* PN j=1 XE 2 j). LS and SS are two-dimensional two-dimensional vectors. To get the center and radius of the microcluster, we need to add the following three components:=LS N=s SSN 2=LS N 2 c=LS N=s SSN (6) Timestamp t represents the time of the microcluster. Last modified (timestamps in CEDAS are called 'energy' microclusters). The word "dense" is determined by a parameter indicating the maximum radius a microcluster can correspond to. This is an important process that relies heavily on data. These are user ACSC and CEDAS parameters, but MDSC recognizes them as such and deletes the confidential values manually changed by detection. Both micelles a and b are considered usable under the following conditions. Distance (acen, bcen) is the center of micelle b, and radius br is the radius of micelle a and b. Macro clusters are created as a series of interoperable high-density micro-clusters. MDSC has two components available on the Internet. Assign newly arrived points to real-time clusters (if they exist). Otherwise, it is sent to the buffer. Buffer noise points can contain new cluster seed or drift signals. Periodically searches the buffer for new clusters (using local adaptive search). The ACSC algorithm uses window models and reconnaissance method groups and inspiration. Windows are single paths and clusters that do not overlap flow blocks are displayed in each window during the same path. In the metaphor, "ants" are a small group of the same ants, and "ant's nest" is a group of comparable ants. Nests formed by group processes are returned to the solution. CEDAS treats fine clusters as nodes in the graph. Giant multiples are interconnected to the ribs and are composed of fine bundles. DenStream uses an online / offline paradigm. After forming the microclusters in it online, DBS is used instead of the traditional clustering method for offline clustering. Example (Cao, 2006)

Classifier Model

Object masks are retained and masks are organised in the conventional manner. Some examples are provided in the next section. When a point is received, a copy of the point is placed in the forwarding offline buffer for the clustering algorithm, which is then used to process the point. When the buffer chosen from the buffer reaches a certain size, the form mask is updated accordingly. This mask is used to group elements until the next point in the sequence is reached. The beta window is the name given to this section of the game. The following is a process for updating and maintaining a dynamic performance mask (DFM), as well as the grouping approach that was utilised to create this mask. Given a window containing d-dimensional points, we first choose the functions that are not considered and then extract the n important features from the window. This subset of functions is referred to as CF. Cfn is written in the following format: Cf1=cf1,., Cfn where cfi=N+d instead of cfi. The Current Mask CM function is a subset of the CF function. A binary mask is created by combining the commands CM=cm1, cm2,.., Cmd. The major CM=d properties are represented as 1 and the remaining properties are expressed as 0. where cmi=(1: possible=CF 0: other digit (8) main CM=d and CFn properties are expressed as 1 and the rest are expressed as 0. CF and CM are computed and updated in each window, and the result is a vector of integers with eigenvalues (FV), where: the meaning or perceived importance of each function at a certain moment in time is: CF and CM FV=fv1,..., Fvd, where fvi is a function of time. Assume that the function fvi is supported. After each window, do an update. In the case of FV with CM value of:=fvi + cmi 2 (9) This is the weighted moving average of the significance of each characteristic (corresponding to the CM of each window) After that, the DFM is modified in accordance with the semantics of the features as follows: There is a PV curve and a threshold that has been set. DFM=dfm1, dfm2, dfm3,..., dfmd where: dfmi=(1: iffvior0: (10) threshold (allfeaturesR 0) to featuresall features iffvior0: (10) threshold (allfeaturesR 0) to featuresall features In the case that no function is selected from the initial set of n functions, this parameter defines the length of the features being examined. A high threshold makes it more difficult to assess new features and more difficult to reject them. While not connected to the internet, the function values advance over time after each. Check. We begin by reading the -points from the buffer and creating a DFM, which we then group according to the mask. The DFM and cluster are formed once the startup process is completed. In DFM, entry points are clustered together. In this case, the microclusters and entrance points of microclusters that are most suited for density clustering methods are used to construct the clusters. Calculate the distance between the point p and the centre c of the microcluster m, keeping in mind that the distance is smaller than the radius r of the microcluster m's centre. The distance between the centre of each fi form and the nearest point on the p scale is measured in units of p. We expect to be published and separated from key characteristics as a result of our work with DFM. The linear sum (LS) and sum of squares (SS) need N points specified by the radius m for the centre c of the microscope and the radius m (Equations (5) and (6)), respectively (SS). As an example, for reconstruction, the function I describes N points (m=1, N), and the instance (j) describes N points (m=1, N). The feature set of m is made up of d features, which are represented by the function (XE j, I=1, N). It is possible to compute the linear sum of characteristic I using the formula LSi=1 NPN j=1 Xij, and the sum of squares using the formula SSi=1 NPN j=1 Xij 2. Apply a mask to each feature and multiply the results.

Algorithm

```
if <cntr mod α== 0> then
Recent Features (RF) Select (buffer) Use RF to generate Recent Mask (RM) Use RM to update Feature Numbers (FN) Use FN to update Feature Mask (DFM) Clear buffer
Store latest FN offline Apply DFM to p
for <every cluster C> do Apply DFM to C Clst(p)
Add copy of p to buffer cntr ++
Read next point
```

RESULT AND DISCUSSION

In each experiment, the proposed DFM method improves the performance of the underlying clustering algorithm applicable to each clustering method evaluated. Three function selectors are used to create the mask. For small image streams, the MCFS method provides the best mask for this situation. Maximum variance gives the best mask in the highest dimensional text stream because the variance is maximum. Since there is no mask, these methods cannot provide results from high-dimensional text streams (up to 60,000 functions). It is more efficient and faster to use static masks (the mask for the function selected in the first window has not changed during the transfer). Dynamic masks, on the other hand, further improve this performance. You can easily track and track the importance of features over time, which is especially useful for statistical analysis. Newsgroup data (Figure 2) is used to illustrate this. We chose two features and used feature drift to track the estimated importance of each feature along a stream. For smaller image streams, you can find a clustering solution without masks, but dynamic masks can improve efficiency. This not only increases speed, but also reduces processing time. Fewer properties are checked during the grouping process. That is, each pair has fewer computations. The grouping efficiency of static masks degrades significantly when applied to image sequences (about 1000 dimensions). It is not recommended to use the function selection method in place of the traditional static selection method, as the concept evolves and yields better results. The static mask is not updated if the new function is duplicated outside the function. It omits major new features that have emerged along unrelated routes. The Lapasian scoring method does not select the features that lead to the best mask, but as shown in Table 8, the maximal distribution method chooses the best features in a larger text stream to complete. Requires less processing time. Where N is a d-dimensional example. , D is the dimension of the instance. MCFS is time consuming and (requires O (N2 + d3) (Cai, 2010) and is suitable for (relatively) small size streams.

CONCLUSION

This paper presents a Dynamic Feature Mask (DFM) for unsupervised dynamic feature selection in non-stationary data streams. Redundant features are masked and clustering is performed along unmasked, relevant features. If a feature’s perceived importance changes, the mask is updated accordingly - previously unimportant features can be unmasked and features which lose relevance become masked. The method is proposed to address two challenges in data stream clustering: 1) feature drift - a change at the feature level in a stream, and 2) the problem of clustering high-dimensional streams where the curse of dimensionality renders distance measurements and the concepts of ‘density’ difficult. The proposed method is algorithm-independent and can be used with any existing density based clustering algorithm There are many density-based clustering algorithms in the literature and they typically do not have a mechanism to deal with feature drift or with very high-dimensionality. We evaluated the proposed method on four density based clustering algorithms (MDSC, CEDAS,ACSC, and DenStream) across four high-dimensional streams; two text streams and two image streams. In each case, the proposed DFM improves clustering performance and furthermore, reduces the processing time required by the underlying algorithm. An unsupervised feature selection method is required to create and maintain the DFM and we evaluate three existing methods: Laplacian Score, Multi-Cluster Feature Selection, and Maximum Variance. Experimental results suggest that on the lower dimensional (≈ 1, 000 dimensions) streams, MCFS is the best selector for the mask. On the higher dimensional text streams (up to 60,000 dimensions), the Maximum Variance method selects the best features to maintain the mask. The Laplacian Score did not return the best features on any stream and was shown to require considerably more time than the other two methods. On each stream, we compare the DFM with a static feature mask. In the static case, the mask is created on one window at the beginning of the stream and is never updated. The dynamic mask performs better on each stream. On the higher dimensional streams, the static mask is preferable to no mask (without a mask the clustering algorithms could not return a solution at all) but on the lower dimensional streams it is preferable to use no mask rather than a static mask. Future work will investigate the suitability of the proposed method for density-based classification methods in high-dimensional data streams with feature drift.

REFERENCES

Bhatnagar, V., Kaur, S., & Chakravarthy, S. (2014). ‘‘Clustering data streams using grid-based synopsis,’’ *Knowl. Inf. Syst.*, *41*(1), 127–152.

Biesiada, J., & Duch, W. (2007). ‘‘Feature selection for high-dimensional data—A pearson redundancy based filter,’’ in *Proc. Comput. Recognit. Syst*. Berlin, Germany: Springer, 242–249.

Bifet, A., & Gavaldà, R. (2009). ‘‘Adaptive learning from evolving data streams,’’in *Proc. Int. Symp. Intell. Data Anal. Berlin, Germany: Springer,* 249–260.

Cai, D., Zhang, C., & He, X. (2010). ‘‘Unsupervised feature selection for multicluster data,’’ in *Proc. 16th ACM SIGKDD Int. Conf. Knowl. DiscoveryData Mining,* 333–342.

Cao, F., Ester, M., Qian, W., & Zhou, A. (2006). ‘‘Density-based clustering over anevolving data stream with noise,’’ in *Proc. SIAM Int. Conf. Data Mining,* 328–339.

Carvalho, V.R., & Cohen, W.W. (2006). ‘‘*Single-pass online learning: Performance, voting schemes and online feature selection,*’’ in Proc. 12th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining, 548–553.

Chandrashekar, G., & Sahin, F. (2014). ‘‘A survey on feature selection methods,’’ *Comput. Elect. Eng., 40*(1), 16–28.

Ester, M., Kriegel, H.P., Sander, J., & Xu, X. (1996). ‘‘A density-based algorithmfor discovering clusters in large spatial databases with noise,’’ in *Proc.KDD, 96*(34), 226–231.

Fahim, A.H., Salem, A.M., Torkey, F.A., & Ramadan, M.A. (2006). ‘‘Densityclustering based on radius of data (DCBRD),’’ *World Acad. Sci., Eng.Technol*.

Fahy, C., & Yang, S. (n.d.). ‘‘*Finding and tracking multi-density clusters in online dynamic data streams,*’’ IEEE Trans. Big Data.

Fahy, C., Yang, S., & Gongora, M. (2019). ‘‘Ant colony stream clustering: A fastdensity clustering algorithm for dynamic data streams,’’ *IEEE Trans.Cybern., 49*(6), 2215–2228.

Fiscus, J., Doddington, G., Garofolo, J., & Martin, A. (1999). ‘‘NIST’s 1998 topicdetection and tracking evaluation (TDT2),’’ in *Proc. DARPA Broadcast News Workshop,* 19–24.

Forestiero, A., Pizzuti, C., & Spezzano, G. (2013). ‘‘A single pass algorithmfor clustering evolving data streams based on swarm intelligence,’’ *Data Mining Knowl. Discovery, 26*(1), 1–26.

Freund, Y., Schapire, R., & Abe, N. (1999). ‘‘A short introduction to boosting,’’*J. Jpn. Soc. Artif. Intell., 14,* 771–780.

Gomes, H.M., Bifet, A., Read, J., Barddal, J.P., Enembreck, F., Pfharinger, B., Holmes, G., & Abdessalem, T. (2017). ‘‘Adaptive random forestsfor evolving data stream classification,’’ *Mach. Learn., 106*, 9–10.

Gu, Q., Li, Z., & Han, J. (2011). ‘‘*Generalized Fisher score for feature selection,*’’in Proc. 27th Conf. Uncertainty Artif. Intell, 266–273.

Guha, S., Mishra, N., Motwani, R., & O’Callaghan, L. (2000). ‘‘Clustering datastreams,’’ in *Proc. 41st Annu. Symp. Found. Comput. Sci. Berlin, Germany:Springer,* 359–366.

Hall, M. (1999). ‘‘Correlation-based feature selection for machine learning,’’Ph.D. dissertation, Dept. Comput. Sci., Waikato Univ., New Zealand,Hamilton.

He, X., Cai, D., & Niyogi, P. (2005). ‘‘*Laplacian score for feature selection,*’’ inProc. 18th Int. Conf. Neural Inf. Process. Syst., 507–514.

Huang, C.L., & Wang, C.J. (2006). ‘‘A GA-based feature selection and parameters optimizationfor support vector machines,’’ *Expert Syst. Appl, 31*(2), 231–240.

Huang, H., Yoo, S., & Kasiviswanathan, S.P. (2015). ‘‘Unsupervised featureselection on data streams,’’ in P*roc. 24th ACM Int. Conf. Inf. Knowl.Manage*. 1031–1040.

Hyde, R., Angelov, P., & MacKenzie, A.R. (2017). ‘‘Fully online clustering of evolving data streams into arbitrarily shaped clusters,’’ *Inf. Sci.,vols*. 382–383.

Jardine, N., & van Rijsbergen, C.J. (1971). ‘‘The use of hierarchic clusteringin information retrieval,’’ *Inf. Storage Retr.,. 7*(5), 217–240.

Katakis, I., Tsoumakas, G., & Vlahavas, I. (2005). ‘‘On the utility of incrementalfeature selection for the classification of textual data streams,’’ in *Proc.Panhellenic Conf. Inform,* 338–348.

Kremer, H., Kranen, P., Jansen, T., Seidl, T., Bifet, A., Holmes, G., & Pfahringer, B. ‘‘*An effective evaluation measure for clustering on evolvingdata streams*,’’ in Proc. 17th ACM SIGKDD Int. Conf. Knowl. DiscoveryData Mining, 868–876.

Law, M.H.C., Figueiredo, M.A.T., & Jain, A.K. (2004). ‘‘*Simultaneous feature selection and clustering using mixture models,*’’ *IEEETrans. Pattern Anal. Mach. Intell, 26*(9), 1154–1166.

Lee, C., & Lee, G.G. (2006). ‘‘Information gain and divergence-based featureselection for machine learning-based text categorization,’’ *Inf. Process. Manage, 42*(1), 155–165.

Li, J., Cheng, K., Wang, S., Morstatter, F., Trevino, R.P., Tang, J., & Liu, H. ‘‘Feature selection: A data perspective,’’ *ACM Comput. Surv, 50*(6), *94*.

Ma, Z., Lai, Y., Kleijn, W.B., Song, Y.Z., Wang, L., & Guo, J. (2019). ‘‘VariationalBayesian learning for Dirichlet process mixture of inverted Dirichlet distributions in non-Gaussian image feature modeling,’’ *IEEE Trans. NeuralNetw. Learn. Syst., 30*(2), 449–463.

Ma, Z., Yu, H., Chen, W., & Guo, J. (2019). ‘‘Short utterance based speech language identification in intelligent vehicles with time-scale modifications and deep bottleneck features,’’ *IEEE Trans. Veh. Technol, 68*(1), 121–128.

Masud, M., Gao, J., Khan, L., Han, J., & Thuraisingham, B.M. (2011). ‘‘Classification and novel class detection in concept-drifting data streams under timeconstraints,’’ *IEEE Trans. Knowl. Data Eng, 23*(6), 859–874.

Nene, S.A., Nayar, S.K., & Murase, H. (1996). ‘‘*Columbia object image library(coil-20),*’’ Columbia Univ., New York, NY, USA, Tech. Rep. CUCS-006-96, 1996.

Nguyen, H.L., Woon, Y.K., Ng, W.K., & Wan, L. (2012). ‘‘Heterogeneousensemble for feature drifts in data streams,’’ in Proc. *Pacific–Asia Conf.Knowl. Discovery Data Mining.* Berlin, Germany, 1–12.

Rand, W.M. (1971). ‘‘Objective criteria for the evaluation of clustering methods,’’ *J. Amer. Stat. Assoc, 66*(336), 846–850.

Roffo, G., Melzi, S., & Cristani, M. (2015). ‘‘Infinite feature selection,’’ in Proc. *IEEE Int. Conf. Comput. Vis. (ICCV),* 4202–4210.

Tu, L., & Chen, Y. (2009). ‘‘Stream data clustering based on grid density andattraction,’’ *ACM Trans. Knowl. Discovery Data, 3*(3), 12.

Wan, L., Ng, W.K., Dang, X.H., Yu, P.S., & Zhang, K. (2009). ‘‘Density-basedclustering of data streams at multiple resolutions,’’ *ACM Trans. Knowl.Discovery Data, 3*(3), 14.

Wang, L., & Shen, H. (2016). “*Improved data streams classification with fastunsupervised feature selection,*’’ in Proc. 17th Int. Conf. Parallel Distrib.Comput., Appl. Technol., 221–226.

Wilcoxon, F., & Wilcox, R.A. (1964). *Some rapid approximate statistical procedures*. Pearl River, NY, USA: Lederle Laboratories, 1964.

Yu, H., Tan, Z.H., Ma, Z., Martin, R., & Guo, J. (2018). ‘‘Spoofing detection inautomatic speaker verification systems using DNN classifiers and dynamicacoustic features,’’ *IEEE Trans. Neural Netw. Learn. Syst, 29*(10), 4633–4644.

Received: 26-Dec-2021, Manuscript No. JMIDS-21-9341; **Editor assigned**: 28-Dec-2021, PreQC No. JMIDS-21-9341(PQ); **Reviewed**: 07-Jan-2022, QC No. JMIDS-21-9341; **Revised**: 19-Jan-2022, Manuscript No. JMIDS-21-9341 (R); **Published**: 26-Jan-2022

Citation Information: Velan, S.G., Somasundaram, K., & Rajavarman, V.N. (2022). An experimental evaluation of clustering and classification of high-speed dimensional data stream in dynamic feature selection. *Journal of management Information and Decision Sciences, 25*(S2), 1-9.