IARJSET
ISSN (Online) 2393-8021
ISSN (Print) 2394-1588
International Advanced Research Journal in Science, Engineering and Technology
ISO 3297:2007 Certified
Vol. 3, Issue 9, September 2016

Efficient Method for Detecting and Localizing Concept drifts from Event logs in Process Mining

Trupti Kakkad[1], Dr. Rahila Sheikh[2]
Rajiv Gandhi College of Engineering and Research Technology Chandrapur, Babupeth, Chandrapur[1, 2]

Abstract: A generic framework and specific techniques to discover once a process changes and to localize the elements of the process that have modified. Totally different options are projected to characterize relationships among activities. These options are wont to discover variations between serial populations. The drift might even be periodic (e.g., due to seasonal influences) or one-of-a-kind (e.g., the results of latest legislation). Projecting the information onto a random lower-dimensional mathematical space yields results love standard spatial property reduction ways like principal part analysis: the similarity of knowledge vectors is preserved well below random projection. Random projections (RP) is computationally significantly more cost-effective than mistreatment, e.g., principal part analysis. RP employing a distributed random matrix provides extra machine savings in random projection.

Keywords: Concept drift, flexibility, hypothesis tests, random projection, dimensionality reduction, image data, text document data, high-dimensional data.

I. INTRODUCTION

Business processes are nothing more than logically connected tasks that use the resources of a company to attain a defined business outcome.
Business processes can be viewed from variety of Perspectives, together with the management flow, data, and also the resource views. In today's dynamic marketplace, it's progressively necessary for enterprises to contour their processes thus on Reduce cost and to enhance performance.

In addition, today's customers expect organizations to be flexible and adapt to ever-changing circumstances. New legislations like the WABO act [1] and also the Sarbanes–Oxley Act [2], extreme variations in offer and demand, seasonal effects, natural calamities and disasters, point in time escalations [3], and so on, also are forcing organizations to vary their processes. For instance, governmental and insurance organizations reduce the fraction of cases being checked once there's an excessive amount of work within the pipeline another example, during a disaster, hospitals, and banks change their operational procedures. it's evident that the economic success of a company is a lot of and a lot of captivated with its ability to react and adapt to changes in its operational environment. Therefore, flexibility and alter are studied in-depth within the context of business method management (BPM). For instance, method-aware data systems (PAISs) [4] has been extended to be able to flexibly adapt to changes in the process. Progressive workflow management (WFM) and metronome marking systems [5] give such flexibility, e.g., we are able to simply release a brand new version of a method. In addition, in processes not driven by WFM/BPM systems (such because the usage of medical systems) there's even a lot of flexibility as processes are controlled by individuals instead of data systems.

Many of today's information systems area unit recording associate degree abundance of event logs. Method mining may be a comparatively young analysis discipline aimed at discovering, monitoring, and up real processes by extracting knowledge from event logs.

In random projection (RP), the first high-dimensional data is projected onto a lower-dimensional topological space employing a random matrix whose columns have unit lengths. RP has been found to be a computationally efficient, however sufficiently correct methodology for spatial property reduction of high-dimensional knowledge sets. Whereas this methodology has attracted various interest, empirical results area unit thin.
RP as a spatial property reduction tool on high-dimensional image and text data sets. In both application areas, random projection is compared to standard spatial property reduction ways.

II. LITERATURE SURVEY

In this section, the basic concepts in process mining and concept drifts in data mining/machine learning.

A. Process Mining:

Process mining serves a bridge between data mining and business process modelling [6]. Business processes leave trails in a very type of knowledge sources (e.g., audit trails, databases, and dealings logs). Process mining aims at discovering, monitoring, and rising real processes by extracting information from event logs recorded by a spread of systems (ranging from detector networks to enterprise data systems). The starting point for process mining is an occasion log that may be a collection of events. That events are related to process instances (often referred to as cases) and square measure described by

IARJSET
ISSN (Online) 2393-8021
ISSN (Print) 2394-1588
International Advanced Research Journal in Science, Engineering and Technology
ISO 3297:2007 Certified
Vol. 3, Issue 9, September 2016

some activity name. The events at intervals a method instance square measure ordered. Therefore, a method instance is usually drawn as a trace over a collection of activities. Additionally, events will have attributes like timestamps, associated resources (e.g., the person death penalty the activity), transactional data (e.g., start, complete, suspend, and so on), and knowledge attributes (e.g., quantity or kind of customer). For a lot of formal definition of event logs employed in process mining, the reader is observed [6]. Event logs like in Fig. square measure fully commonplace within the process mining community and event log formats like MXML [7] and XES [8] square measure used. The topics in process mining are generally classified into three categories:

1) Discovery;
2) Conformance; and
3) Enhancement.

Process discovery deals with the discovery of models from event logs. These models might describe management flow, structure aspects, time aspects, and so on. Fig. shows the fundamental plan of method discovery. An event log containing elaborate information concerning events is transformed into a multiset of traces L = [abcdjkln, aefjkmn, abgchdjkln...]. Process discovery techniques are able to discover process models like the Petri internet shown in Fig. conformity deals with scrutiny associate a priori method model with the discovered behavior as recorded within the log and aims at detecting inconsistencies/deviations between a process model and its corresponding execution log.

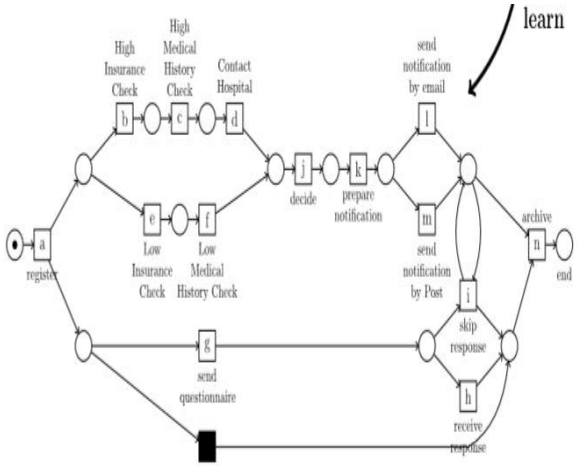

Fig 1: Process Delivery.

B. Concept Drift:

Concept drift [12] in machine learning and data mining refers to things when the relation between the computer file and the target variable, that the model is making an attempt to predict, changes over time in unforeseen ways that. Therefore, the accuracy of the predictions could degrade over time. To forestall that, prognostic models have to be compelled to be able to adapt on-line, i.e., to update themselves frequently with new information. The setting is often coiled over Associate in Nursing infinite data stream as follows: 1) receive new data; 2) make a prediction; 3) receive feedback (the true target value); and 4) update the predictive model. whereas operational under such circumstances, prognostic models area unit required: 1) to react to concept drift (and adapt if needed) as soon as possible; 2) to distinguish drifts from once-off noise and adapt to changes, however be robust to noise; and 3) to control in but information point and use restricted memory for storage. During this setting, several reconciling algorithms are developed.

C. Random Projection:

RP as a spatiality reduction tool on high-dimensional image and text knowledge sets. In each application areas, random projection is compared to documented dimensionality reduction methods. They show that despite the computational simplicity of random projection, it doesn't introduce a significant distortion within the knowledge. The info sets utilized in this paper square measure of terribly different natures. Their image knowledge is from monochrome pictures of natural scenes. A picture is conferred as a matrix of element brightness values, the distribution of that is usually just about Gaussian: symmetric and bell-shaped. Text document knowledge is conferred in vector space, during which every document forms one d-dimensional vector wherever d is that the vocabulary size. The i-th component of the vector indicates (some perform of) the frequency of the i-th vocabulary term within the document. Document knowledge is usually extremely thin or peaked: just some terms from the vocabulary square measure gift in one document, and most entries of the document vector square measure zero. Also, document knowledge contains a no symmetrical, absolutely skewed distribution, because the term frequencies square measure plus. It's instructive to visualize however random projection works as a spatiality reduction tool within the context of those two very different application areas.

Random projection within the pre-processing of matter knowledge, before applying LSI. They gift experimental results on Associate in nursing artificially generated set of documents.

The problems of dimensionality reduction and similarity search have usually been self-addressed within the information retrieval literature and other approaches than random projection is conferred.

III.PROPOSED APPROACH FRAMEWORK

3.1. Architecture

Over the last two decades several researchers are acting on method flexibility. Ploesser et al. [32] have classified business method changes into 3 broad categories: 1) sudden; 2) anticipatory; and 3) evolutionary. This classification is employed during this paper, however currently within the context of event logs.

This approach uses process mining to produce associate aggregative summary of all changes that have happened to date. This approach, however, assumes that modification

IARJSET
ISSN (Online) 2393-8021
ISSN (Print) 2394-1588
International Advanced Research Journal in Science, Engineering and Technology
ISO 3297:2007 Certified
Vol. 3, Issue 9, September 2016

logs square measure obtainable, i.e., modifications of the workflow model are recorded. At now of time, only a few information systems offer such modification logs.
Concept drift is in varied branches of the data mining and machine learning community. Construct drift has been two sorts supervised and unsupervised settings and has been shown to be necessary in several applications. in contrast to in data processing and machine learning, wherever construct drift focuses on changes in straightforward structures like variables, construct drift in method mining deals with changes to complicated artifacts like method models describing concurrency, choices, loops, and cancelation. They work differs from in several ways: 1) this approach constructs associate abstract illustration of a method in contrast to ours wherever we have a tendency to think about options characterizing the traces and 2) this system is applicable just for modification detection whereas our framework is applicable for each modification (point) detection and alter localization.

3.2 Process Flow:
The various aspects of process modification. Initially, they describe modification views (control flow, data, and resource). Then, the various types of drift (sudden, gradual, recurring, periodic, and incremental).

A. Views of Modification: There square measure three necessary perspectives within the context of business processes:
1) management flow;
2) data; and
3) resource.

One or a lot of those perspectives could modification over time.
1) Control flow/behavioral perspective: This category of changes deals with the behavioral and structural changes in an exceedingly method model. Rather like the look patterns in computer code engineering, there exist modification patterns capturing the common control-flow changes.
Control flow changes is classified into operations like insertion, deletion, substitution, and reordering of process fragments.
2) Data perspective: This category of changes visits the changes within the production and consumption of data and therefore the result of data on the routing of cases. As an example, it should now not be needed to possess a specific document once approving a claim.
3) Resource perspective: This category deals with the changes in resources, their roles, and structure, and their influence on the execution of a process. As example, certain execution methods in an exceedingly process may well be enabled (disabled) upon the supply (non availability) of resources.

B. Nature of Drifts: With the duration that a modification is active, they can classify changes into momentousness and permanent. momentousness changes square measure short lived and have an effect on only a very few cases, whereas permanent changes square measure persistent and stay for a moment, they concentrate on permanent changes as momentary changes usually can't be discovered owing to insufficient data.

1) Sudden drift: This corresponds to a substitution of an existing process M1 with a new process M2, this class of drifts is typically seen in scenarios such as emergencies, crisis situations, and change of law. As an example, a new regulation by the finance ministry of India mandates all banks to procure and report the customer's personal account number in their transactions.
2) Gradual drift: Unlike the sudden drift, here both processes coexist for some time with M1 discontinued gradually.For example, a supply chain organization might introduce a new delivery process. This process is, however,applicableonly fororders takenhenceforth.All previous orders still have to follow the former delivery process.
3) Recurring drift: It is quite natural to observe such a phenomenon with processes having a seasonal influence. For example, a travel agency might deploy a different process to attract customers during Christmas period. The recurrence of processes may be periodic or non-periodic.An example of a non-periodic recurrence is the deployment of a process subjected to market conditions.
4) Incremental drift: This class of drifts is more pronounced in organizations adopting an agile BPM methodology and in processes undergoing sequences of quality improvements (most total quality management) initiatives are examples of incremental change

3.3 Random Projection
In random projection, the original d-dimensional data is projected to a k-dimensional (k << d) subspace through the origin, using a random k ×d matrix R whose columns have unit lengths. Using matrix notation where X_n is the original set of N d-dimensional observations,

$$X^{Rp}_{k\times N} = R_{k}\times_{d} X_{d}\times_{N}$$

Is the projection of the data onto a lower k-dimensional subspace? The key idea of random mapping arises from the Johnson-Linden Strauss lemma [15]: if points in a vector space are projected onto a randomly selected subspace of suitably high dimension, then the distances between the points are approximately preserved.
Random projection is computationally very simple: forming the random matrix R and projecting the d × N data matrix X into k dimensions is of order O(dkN), and if the data matrix X is sparse with about c nonzero entries per column, the complexity is of order O(ckN).

3.4 Discrete cosine transform (DCT)
Discrete cosine transform (DCT) may be a wide used methodology for compression and as such it can also be employed in spatial property reduction of image data. DCT is computationally less burdensome than PCA and its performance approaches that of PCA. DCT is additionally best for human eye: the distortions introduced occur at the

IARJSET
ISSN (Online) 2393-8021
ISSN (Print) 2394-1588
International Advanced Research Journal in Science, Engineering and Technology
ISO 3297:2007 Certified
Vol. 3, Issue 9, September 2016

best frequencies only, and the human eye tends to neglect these as noise. DCT is performed by easy matrix operations a picture is remodeled to the DCT house and spatial property reduction is finished within the inverse remodel by discarding the remodel coefficients corresponding to the best frequencies. Computing the DCT isn't data-dependent, in distinction to PCA that wants the eigenvalue decomposition of information variance matrix; that's why DCT is orders of magnitude cheaper to cypher than PCA. Its process complexness is of the order O (dN log2 (dN)) for a knowledge matrix of size d×N.

3.5 Mathematical Model:

a)In random projection, the original d-dimensional data is projected to a k-dimensional (k << d)

$$X^{Rp}_{k\times N} = R_{k}\times_{d} X_{d}\times_{N}$$

Using matrix notation where Xd×N is the original set of N d-dimensional.

b) Euclidean distance between two data vectors x1 and x2 in the original large-dimensional space as ||x1−x2||.

C) After the random projection, this distance is approximated by the scaled Euclidean distance of these vectors in the reduced space:

$\sqrt{d/k}\,\|R_{x1}-R_{x2}\|$

where d is the original and k the reduced dimensionality of the data set. The scaling term $\sqrt{d/k}$ takes into account the decrease in the dimensionality of the data.

D) The choice of the random matrix R is one of the key points of interest. The elements rij of Rare often Gaussian distributed, but this need not be the case.

$$r_{ij}=\sqrt{3}\begin{cases}+1 \text{ With probablity } \frac{1}{6}\\ 0 \text{ With Probablity } \frac{2}{3}\\ -1 \text{ With Probablity } \frac{1}{6}\end{cases}$$

E) If dimensionality reduction of the data set is desired; the data can be projected onto a subspace spanned by the most important eigenvectors:

$$X^{PCA}=E_k^T X$$

where the d×k matrix Ek contains the k eigenvectors corresponding to the k largest eigenvalues.

F) A closely related method is singular value decomposition (SVD): $X = USV^{T}$ where orthogonal matrices U and V contain the left and right singular vectors of X, respectively, and the diagonal of S contains the singular values of X.

$$X^{SVD}=U_k^T X$$

where U_k is of size d×k and contains these k singular vectors. Like PCA, SVD is also expensive to compute. S Letter templates for Latex and Microsoft Word. The Latex templates depend on the officialIEEEtran.cls and IEEEtran.bst files, whereas the Microsoft Word templatesare self-contained.

IV. FEATURE EXTRACTION

Event logs are characterized by the relationships between activities. Dependencies between activities in an event log can be captured and expressed using the follows (or precedes) relationship, also referred to as causal footprints. For any pair of activities, a, b ∈A, and a trace t = t (1) t (2) t (3) ...t(n) ∈ A+, we say b follows an if and only if for all $1 \le i \le n$ such that
t (i) = a there exists a j such that $i < j \le n$ and t (j) = b. In temporal logic notation: (a ⇒ (♦b)).

They distinguish between two classes of features:
1) Global and
2) Local features.

Global features are defined over an event log, whereas local features can be defined at a trace level. With the follows (precedes) relation, they propose two global features:
1) Relation type count (RC) and
2) Relation entropy (RE), and two local features:
a) Window count (WC) and
b) J measure.

1) RC:
The RC with respect to the follows (precedes) relation is a function,

$$f^{L}_{Rc}: A\rightarrow N_0\times N_0\times N_0$$

$f^{L}_{Rc}$of an activity, x ∈A, with respect to the follows (precedes) relation over an event log L is the triple {cA, cS, cN} where {cA, cS, and cN} are the number of activities in A that always, sometimes, and never follows (precedes) x, respectively, in the event log L.

2) RE:
The RE with respect to the follows (precedes) relation is a function,

$$f^{L}_{RE}: A\rightarrow R+0,$$

Defined over the set of activities $f^{L}_{RE}$of an activity, x ∈A with respect to the follows (precedes) relation is the entropy of the RC metric.

3) WC:
Given a window of size l ∈ N, the WC with respect to follows (precedes) relation is a function,

$$f^{L}_{RC}: A\times A\rightarrow N0,$$

defined over the set of activity pairs. Given a trace t and a window of size l, letS^{l}(a) be the bag of all subsequence'st (i, i +l−1), such that t(i)=a.[3]

4) J Measure:
J measure with respect to follows (precedes) relation is a function

IARJSET
ISSN (Online) 2393-8021
ISSN (Print) 2394-1588
International Advanced Research Journal in Science, Engineering and Technology
ISO 3297:2007 Certified
Vol. 3, Issue 9, September 2016

$$f^{l,t}_{j}: A \times A \rightarrow R+$$

Defined over the set of activity pairs and a given window of length $l \in N$. Let pt.(a) and pt.(b) are the probabilities of occurrence of activities an and b, respectively, in a trace **t**.

V. FEATURE EXTRACTION

They propose the framework shown in Fig. for analysing concept drifts in process mining. The framework identifies the subsequent steps:

1) Feature extraction and selection: This step pertains in defining the characteristics of the traces in an incident log. During this paper, they need outlined four options that characterize the control-flow perspective of process instances in an incident log. Looking on the main focus of research, they need define further options.

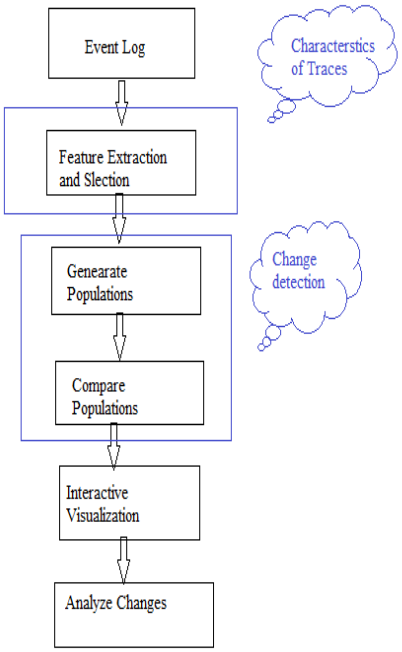

Figure 2: System Architecture

2) Generate populations: an incident log may be remodelled into an information stream supported the options designated within the previous step. This step deals with defining the sample populations for learning the changes within the characteristics of traces. Completely different criteria/scenarios are also thought of for generating these populations from the data stream.

3) Compare populations: Once the sample populations are generated, the next step is to analyse these populations for any change in characteristics. They advocate the utilization of applied mathematics hypothesis tests for comparison populations. The null hypothesis in applied mathematics tests states that distributions (or suggests that, or commonplace deviations) of the two sample populations are equal. Looking on desired assumptions and also the focus of research, completely different applied mathematics tests may be used.

4) Interactive visualization: The results of comparative studies on the populations of trace characteristics may be intuitively given to Associate in nursing analyst. As an example, the significance chances of the hypothesis tests may be unreal as a drift plot. Troughs in such a drift plot signify an amendment within the significance likelihood thereby implying an amendment within the characteristics of traces.

5) Analyse changes: image techniques like the drift plot will assist in distinguishing the change points. Having identified that a change had taken place, this step deals with techniques that assist Associate in Nursing analyst in characterizing and localizing the change and in discovering the change process.

VI. FEATURE EXTRACTION

In this section we are discussing the practical environment, scenarios, performance metrics used etc.

6.1 Input:
In this Training and Testing Image is the input for our practical experiment.

6.2 Hardware Requirements:

- RAM : - 1GB
- Processor : -P-IV– 500 MHz to 3.0 GHz
- Disk : -20 GB
- Monitor : -Any Color Display
- Standard Keyboard and Mouse

6.3 Software Requirements:

- Operating System : -Windows 7/XP
- Programming Languages : - Java
- Database Server : - My Sql
- Web server : - Apache Tomcat

6.4 Results of Practical Work**:**
Following figures are showing results for practical work which is done.

Following figure showing the main screen. That takes the input data set,

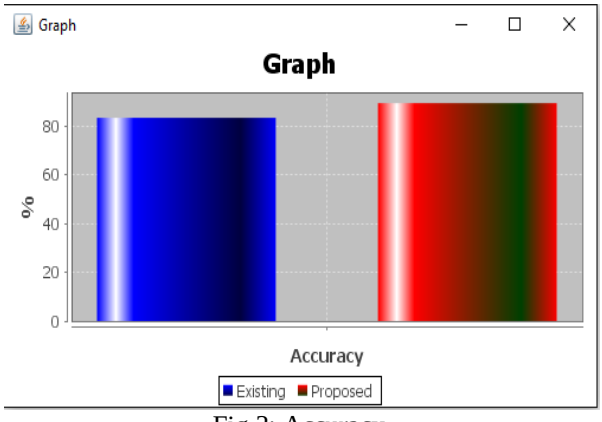

Fig 3: Accuracy

IARJSET
ISSN (Online) 2393-8021
ISSN (Print) 2394-1588
International Advanced Research Journal in Science, Engineering and Technology
ISO 3297:2007 Certified
Vol. 3, Issue 9, September 2016

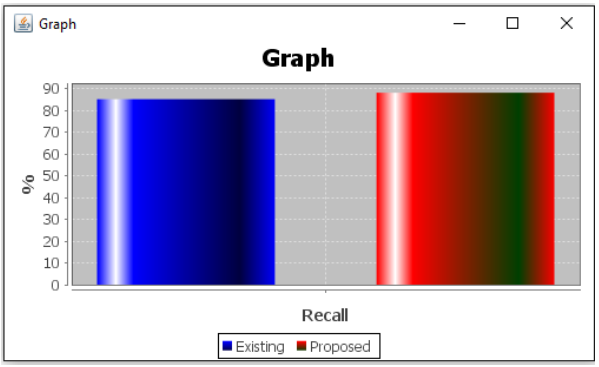

Fig 4: Recall

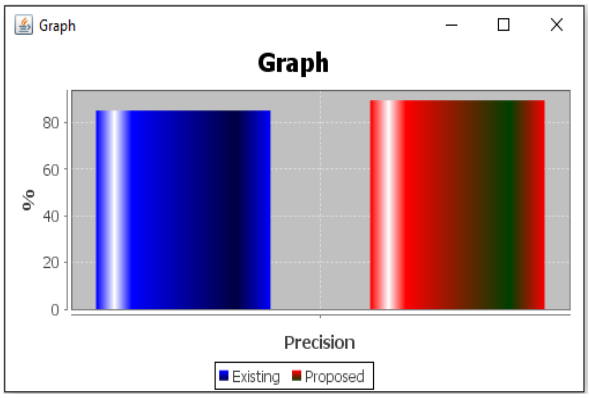

Fig 5: Precision

VII. CONCLUSION AND FUTURE WORK

We have studied the topic of concept drift in process mining and random projection i.e., analysing process changes based on event logs. They projected feature sets and techniques to effectively notice the modifications in event logs and determine the regions of change during a method. Their initial results show that heterogeneousness of cases arising because of process changes is effectively prohibited by police investigation idea drifts. Random projection in spatiality reduction of high-dimensional real-world information sets. Once scrutiny completely different ways for spatiality reduction, the standards are the number of distortion caused by the strategy and its procedure complexness. Their results indicate that random projection preserves the similarities of the data vectors well even when the data is projected to moderate numbers of dimensions; the projection is yet fast to compute. They conclude that random projection is a good alternative to traditional, statistically optimal methods of dimensionality reduction that are computationally infeasible for high dimensional data. Random projection does not suffer from the curse of dimensionality, quite contrary to the traditional methods.

REFERENCES

[1] Ella Bingham and Heikki Mannila "Random projection in dimensionality reduction: Applications to image and text data".
[2] D. Achlioptas. Database-friendly random projections. In Proc. ACM Symp. on the Principles of Database Systems, pages 274–281, 2001.
[3] C. C. Aggarwal, J. L. Wolf, and P. S. Yu. A new method for similarity indexing of market basket data. In Proc. 1999 ACM SIGMOD Int. Conf. on Management of data, pages 407–418, 19993) G. Salton and M.J. McGill. Introduction to modern information retrieval. McGraw-Hill, 1983.
[4] R. Agrawal, C. Faloutsos, and A. Swami. Efficient similarity search in sequence databases. In Proc. 4th Int. Conf. of Data Organization and Algorithms, pages 69–84. Springer, 1993.
[5] R. I. Arriaga and S. Vempala. An algorithmic theory of learning: robust concepts and random projection. In Proc. 40th Annual Symp. on Foundations of Computer Science, pages 616–623. IEEE Computer Society Press, 1999.
[6] M.-W. Berry. Large-scale sparse singular value computations. International Journal of Super-Computer Applications, 6(1):13–49, 1992.
[7] M.-W. Berry. Large-scale sparse singular value computations. International Journal of Super-Computer Applications, 6(1):13–49, 1992.
[8] S. Dasgupta. Experiments with random projection. In Proc. Uncertainty in Artificial Intelligence, 2000.
[9] W.B. Johnson and J. Lindenstrauss. Extensions of Lipshitz mapping into Hilbert space. In Conference in modern analysis and probability, volume 26 of Contemporary Mathematics, pages 189–206. Amer. Math. Soc., 1984.
[10] G. Salton and M.J. McGill. Introduction to modern information retrieval. McGraw-Hill, 1983.
[11] R. P. Jagadeesh Chandra Bose, Wil M. P. van der Aalst, Indre˙ Žliobait˙e, and Mykola Pechenizkiy "Dealing with Concept Drifts in Process Mining" IEEE transactions on neural networks and learning systems, vol. 25, no. 1, january 2014.
[12] H. Schonenberg, R. Mans, N. Russell, N. Mulyar, and W. M. P. van der Aalst, "Process flexibility: A survey of contemporary approaches," in Proc. Adv. Enterprise Eng. I, 2008, pp. 16–30.
[13] M. Sonka, V. Hlavac, and R. Boyle. Image processing, analysis, and machine vision. PWS Publishing, 1998.

BIOGRAPHY

Trupti Kakkad receives her Bachelor in engineering degree in Information Technology from Rajiv Gandhi College of Engineering and Research Technology, Chandrapur in 2007 - 2008. She has a work experience of 5 years in TATA Consultancy Services.