International Journal on Recent and Innovation Trends in Computing and Communication ISSN: 2321-8169
Volume: 5 Issue: 3 109 – 114

# Multi-Label Classification Using Noise Reduction Technique

Ms. Shrutika Yawale
Dept. of InformationTechnology,
MIT College of Engineering.
Pune
*shrutikayawale04@gmail.com*

Prof. Vaishali Suryawanshi.
Dept. of Information Technology,
MIT College of Engineering.
Pune
*vaishali.suryawanshi@mitcoe.edu.in*

***Abstract*—**In domain of data mining and machine learning, multi-label classification is widely studied research problem. The goal of multi-label classification is to predict the absence or presence certain labels of a particular applications those are associated with different classes. In this paper, IML-Forest method is presentedwith goal of improving the performance of multi-label classification over different types of datasets. IML-Forest is based on existing ML-Forest technique. In this paper the construction of set of hierarchical trees and designed the label transfer mechanism in order to identify multiple relevant labels in hierarchical way is proposedto solve the problem of label dependencies in multi label classification. Basically relevant labels at higher levels of trees capture the more discriminable label concepts; next they will be shifted at lower level nodes. From the hierarchy the relevant labels are further aggregated in order to compute the label dependency and make the classification prediction. The problem with ML-Forest method is that noise considerations not yet addressed as collected multi-label dataset may be noisy and imbalanced. This can degrade the performance of learning and accuracy. Noise reduction method is proposed on multi-label dataset to solve the problem of noisy and imbalanced dataset. In this paper the text noises related to low-level data errors are handled.

***Keywords*—*Multi-label classification; noise; IML Forest; ML Forest;***

*****

## I. INTRODUCTION

To predict thepresence or absence of certain labels of an example which is associated with multiple classes is the main aim of MULTI-LABEL classification. Different from classical multi-class problems, where an example is associated with only one single label, the multi-label classification is more general since real-world objects often contain multiple semantic objects. For example, a real-world image usually belongs to multiple categories based on different context, such as water, ship, etc.; while a text document can be classified into a set of topics, such as news, sports, etc [1]. In the last decades, multi-label classification problem has received broad attention from various research domains, such as text categorization, bioinformatics , and computer vision[1].

The past decade has seen a wide variety of papers published on multi-label document classification, in which each document can be assigned to one or more classes. Start by discussing the limitations of existing multi-label document classificationtechniques when applied to datasets with statistical attributes common to real world problems,such as large numbers of labels with power-law-like frequency statistics is present[2]. Then prompt the application of generative probabilistic models in this context. How these models can be better in the situation of large-scale multi-labelcorpora, through particularly assigning individual words to certain labels within eachdocument—rather than assuming that all of the words within a document are pertinent toeach of its labels, and jointly modelling all labels within a corpus simultaneously, whichlends itself well to the task of accounting for the dependencies between these labels [3] is illustrated in this paper.

The need for augmenting unstructured data with metadata is also increasingwith the increasing amount of textual data on the web and in digital libraries. Extraction of different type ofinformation from unstructured text, from minor information such as title and author, to important information such as descriptive keywords and categories is required forsystematically maintaining a high quality digital library [5]. From ever-growing document collections, a non-automatic time and cost-wise extraction of such information is inappropriate.In the literature, one can find a number of multi-label classificationapproaches for a variety of tasks in different domains such as bioinformatics [1], music[2], and text [2]. In the simpler words,a setof binary classification tasks that decides for each label independently whether it shouldbe assigned to the document or notcan be considered as multi-label classification. However, for ongoing research in multi-label classification to focus on the question of how such dependencies can be harnessed,this binary relevanceapproachdoes not consider dependencies between the labels.BP-MLL is one such approach, which formulates multi-label classification problems as a neural network with multiple output nodes, one for each label [4]. The outputlayer is able to model dependencies between the individual labels.

Establishedsingle-labelclassification is anxious with learning from a set of examples that are compatible with a single label from a set of disjoint labels L, |L|> 1. The learningproblem is called abinaryclassification problemif |L|= 2, while if |L|> 2, then it is called a multi-

classclassification problem. The instances are related with a set of labels Y ⊆Linmulti-labelclassification[6]. The tasks of text categorization and medical diagnosismainly motivated multi-label classification. Text documents usually consist of more than one conceptual class. Nowadays, modern applications, such as protein function classification, music categorization and semanticscene classification frequently requires multi-label classification methods. A photograph can correspond to more than one conceptual class, such as sunsetsand beachesat the same time in semantic scene classification. Similarly, a song can correspond to more than one genre in music categorization [7]. For example, tracks of the well-known rock band Scorpionscan be distinguished as both rockand ballad.

In this paper, anew tree ensemblealgorithm, called Ml-Forest is proposed to clearly utilize the label dependency for multi-label classification. In ML-FOREST, a set of hierarchical trees are constructed to learn the label dependency, and then combined as an ensemble to do multi-label prediction. To find a good hierarchical structure so that two relevant instances with strong label dependency will be located in the same node of the tree [8] is the main objective of this paper. To achieve this, a new tree generation algorithm is designed to partition the learning data into smaller subsets from the root to the leaves, and then identify relevant labels for each node with a label transfer mechanism. For the first task of the algorithm, train multi-class classifiers at each node to divide the data into child nodes. Here, each data instance is partitioned into one child node accordingtothe classifierpredictionresults,andtheclasslabel with highest probability given at the node is considered as its relevant label. For the second task of the algorithm, a label transfer mechanismisinvolvedtorecursively propagate therelevant labels from the root down to the leaf node. In the end,each leaf node is characterized by multiple relevant labels given by the nodes at different levels of the tree[3]. This results in a new label dependency portrayal, where the learning models at different levels work jointly and effectively to disclose multiple label concepts belonging to the given data. Intuitively, the relevant labels at high levels in the hierarchy may tend to capture “more significant” label concepts and hence are thematically more general, while the relevant labels at low levels would capture “less significant” label concepts and hence are thematically more specific [5].

The noise can be the difference between the coded representation of the data and the correct, or original data. It can be due to some typing mistakes or colloquialisms always present in natural language and usually reduces the quality of data in a way that makes the data less usable to automated processing by computers such as natural language processing. The noise can also be generated through an with-drawl process (i.e. transcription, OCR) from media other than original electronic texts.

Various business experts state that unstructured data comprise around 80% of the totalenterprise data. A great amount of this data comprises chat transcripts, emails and other informal and semi-formal internal and external communications. Generally such text is meant for human utilization, but - given the amount of data –non-automated processing and appraisal of those resources is not practically sensible anymore. This raises the need for robust text mining methods.

Following are the major contributions of this paper.

A new hierarchical tree algorithm, called IML-TREE with noise removing technique is proposed in this paper, to solve the multi-label classification task. UnliketheBR method whichtransformsthedatainto independentbinaryproblems,ouralgorithmexploits the intrinsic label dependency of the data and incorporates the ML-TREE structure to find the relevant labels of an instance with multiple labels. Hence,aproperwayfor modelling the inherent label dependency of the data into a tree structure is provided by theproposedapproach. A label transfer mechanism is designed to find the relevantlabelsinthehierarchy.Thelabelsofthehigh levels in the hierarchy will be used as priors for the nodes in the low levels to reduce the label space. Therefore,buildingthe classifier modelforlowlevels can be very efficient.

An ensemble strategy is developed to construct multiple hierarchical multi-label trees and combine the predictions of different trees as an ensemble to make predictions.

In this paper the empirical performance is evaluated by conducting an extensive set of experiments on real-world problems in text classification, computer vision and bioinformatics.

## II. LITERATURE SURVEY

In [1], author Timothy N. Rubinexplores a class of productive statisticaltopic models for multi-label documents that connect particular word tokens with distinct labels. Author investigates the advantages of this approach relative to discriminative models, particularly with respect to classification problems involving large numbers of relatively rare labels. Author compare the performance of generative and discriminative approaches on document labelling tasks ranging from datasets with several thousand labels to datasets with tensof labels.

In [2], J. Nam, J. Kim, E. L. Menc´ıa, I. Gurevychinvestigate limitations of BP-MLL, a neural network (NN) architecture that aims at minimizing pair-wise ranking error. Alternatively,they have proposed to use a comparatively simple NN technique with recently

proposedlearning techniques for large-scale multi-label text classification tasks.

In [3], F. Sun, J. Tangproposes to learn a sparse structure of label dependency. The underlying philosophy is that as long as the multi-label dependency cannot be well explained, the principle of parsimony should be applied to the modeling process of the label correlations.

In [4],G. Tsoumakas and I. Katakisperforms comparative experimental results of certain multi-label classification methods andintroduce the task of multi-label classification, organizes the sparse related literature into a structured presentation. It also provides the definition of concepts for the quantification of the multi-label nature of a data set.

In [5],S. Huang, Y. Yu, and Z. Zhou, propose the MAHR approach, which is ableto automatically discover and exploit label relationship. If two labels are related, the hypothesisgenerated for one label can be helpful for the other label, this is their basic idea.A boosting approach with ahypothesis reusemechanism is implemented as idea by MAHR.

## III. PROPOSED APPROACH FRAMEWORK AND DESIGN

### *A. Problem Definition*

In data mining and machine learning domain, the concept of multi-label classification is widely studied research problem. The goal of multi-label classification is to predict the absence or presence certain labels of a particular example those are associated with different classes. As the real world objects are having multiple semantic objects, multi label classification is more general. There are number of methods previously proposed for solving the multi-label classification such as binary relevance in which problem is decomposed into the set of single label multi class problems. The other proposed method then tried to exploit the multiple labels dependencies but effectively modeling of label dependency explicitly is major research problem. Further to solve this problem some more methods introduced in which label dependency learning is conducted from limited information. Over-fitting issue is the difficulty with such methods.ML-Forest method is presented in which new tree ensemble method is applied in order to clearly utilize label dependency for the problem of multi-label classification to conquer these limitations. However, there are several ways to extend the work of ML-Forest method.

### *B. Proposed System Architecture*

In multi-label classification research problems, labels are frequently depends on another labels and hence exploiting the label dependencies is resulted into the accuracy improvement in multi-label classifications. There are two research problems studied in this paper such as efficient label exploiting and noise removal approach. In this paper, IML-Forest method is proposed with goal of improving the performance of multi-label classification over different types of datasets. IML-Forest is based on existing ML-Forest technique. In this paper the construction of set of hierarchical trees and designed the label transfer mechanism in order to recognize multiple relevant labels in hierarchical way is proposedto solve the problem of label dependencies in multi label classification. Basically relevant labels at higher levels of trees capture the more discriminable label concepts, further they will be shifted at lower level nodes. The problem with ML-Forest method is that noise considerations not yet addressed as collected multi-label dataset may be noisy and imbalanced. This can degrade the performance of learning and accuracy. In this paper noise reduction method is proposed on multi-label dataset to solve the problem of noisy and imbalanced dataset. For noise removal, use of hyper clique-based data cleaner method is proposed in this paper.

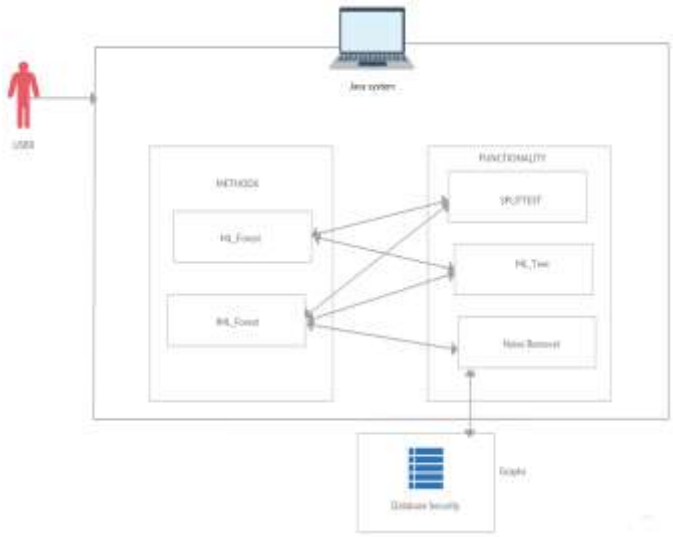

Figure.1. System Architecture

## IV. MATHEMATICAL MODULE

**Algorithm 1 ML-TREE**

**Input:** A training data set D, and a relevant label vector b=none

**Output:** A hierarchical multi-label tree

**Step1 :**(b, h, P) = SPLITTEST(D; b)
**Step 2 :if** h 6= none ^ Acceptable(P) **then**
**Step 3 :for** Di 2 P **do**
**Step 4:** treei=ML-TREE(Di, b)
**Step 5: end for**
**Step 6: return** node(h, b, [iftreeig)
**Step 7: else**
**Step 8: return** leaf(h, b)
**Step 9: end if**

**Algorithm 2 SPLITTEST**
**Input:** A training data set D, a relevant label vector bp from parent
**Output:** A classifier h, a new relevant label vector b, and a partition P for current node

**Step 1**: compute p using Eq. (2)
**Step 2**: calculate b using Eq. (3) and (4)
**Step 3**: (h, P) = (none, ∅)
**Step 4**: h = build classifier on D for those labels which have notbeen identified according to b
**Step 5**: **if** h 6=none **then**
**Step 6**: P= partition D using h
**Step 7**: **end if**
**Step 8**: **return** (b, h, P)

**Algorithm 3 ML-FOREST**

**Training Phase**

**Input:** A training data set D, the number of trees K

**Output:** A forest of tree classifiers F

**Step 1**: F = ∅
**Step 2**: **for** i = 1 to K **do**
**Step 3**: prepare the training set Di = bootstrap(D)
**Step 4**: build tree classifier Ti = ML-TREE(D, none)
**Step 5**: F = F∪Ti
**Step 6**: **end for**
**Step 7**: **return** F

**Classification Phase**
1: For a given x, let b1; _ _ _ ;bK be the predictions assignedby the classifiers, calculate the confidence for each class cj by the average combination method:

$$c^j = \frac{1}{k}\sum_{k=1}^{k} b_k^j$$

2: To the classes with the confidences higher thana predefined threshold value assign x.
For examining the conditional label dependence,the joint conditional probability distribution p(yjx), which defines the probability of the label combination for a particular instance, provides a suitable point of departure. Mathematically, p(y/x) can be written as:

$$\begin{aligned} p(\mathbf{y}|\mathbf{x}) &= p(y^1|\mathbf{x})p(y^2,\cdots,y^q|y^1,\mathbf{x}) \\ &= p(y^1|\mathbf{x})p(y^2|y^1,\mathbf{x})p(y^3,\cdots,y^q|y^1,y^2,\mathbf{x}) \\ &= p(y^1|\mathbf{x})p(y^2|y^1,\mathbf{x})\cdots p(y^q|y^1,\cdots,y^{q-1},\mathbf{x}) \end{aligned}$$

(1)

to find the relevant labels for each node, design a label purity vector, denoted by p = [p1; _ _ _ ;pq]T, to represent the purities of different classes. Specifically, calculate each class label's data purity by

$$p^j = \frac{1}{|D|}\sum_{x_i \in D} y_i^j \qquad (2)$$

where pj 2 [0; 1] is the purity for the j-th class label, D is the examples at the node, and jDj is the number of examples in D. Then build a pertinent label vector, b = [b1; _ _ _ ;bq]>, and integrate the purities into its calculation to find the majority labels as the relevant labels of a node.

$$b^j = \begin{cases} 1, & if p^j \geq \lambda \\ 0, & otherwise \end{cases} \qquad (3)$$

where bj is the relevant label indicator for the j-th classlabel, _ 2 (0:5; 1:0) is a purity threshold.
Our idea is to preserve the identified relevant label vector bp = [b1 p; _ _ _ ;bqp]> from the parent node and incorporate it as an additional indicator with the relevant label vector bc = [b1c ; _ _ _ ; bqc ]> of a child node, which can be obtained a final result of relevant labels b as follows:

$$b^j = \begin{cases} 1, & if b_p^j = 1 \; or b_c^j = 1 \\ x, & otherwise \end{cases} \qquad (4)$$

Basing on the relevance label vectors (i.e., b1; _ _ _ ;bK) from the leaves w.r.t. all K trees, compute the ensemble confidence outputs c by

$$c^j = \frac{1}{k}\sum_{k=1}^{k} b_k^j \qquad (5)$$

where bj k is the j-th element of the relevant label vector bk.
For a testing example x, ML-FOREST outputs a prediction vector y = [y1; _ _ _ ; yq]> with yj = 1 indicating the j-th label is relevant regarding x. Consider a confidence vector c = [c1; _ _ _ ; cq]> 2 Rq for x, where each element of c belongs to a confidence value for one class label. Given w, the prediction y of x can be completed by finding a bipartition of relevant and irrelevant labels based on a threshold function ft(w) such that

$$y^j = \begin{cases} 1, & if w^j \geq t \\ 0, & otherwise \end{cases} \qquad (6)$$

where t 2 [0; 1] is a predefined threshold value. There are several ways to set the threshold value t. For example, set t = 0:5 for simplicity.

**NP-Complete:**

If:

1. *C* is in NP, and
2. Every problem in NP is reducible to *C* in polynomial time.

A decision problem *C* is NP-complete.

*C* can be shown to be in NP by demonstrating that a candidate solution to *C* can be verified in polynomial time.A problem fulfilling condition 2 is said to be NP-hard, whether or not it fulfill condition 1.A consequence of this definition is that if had a polynomial time algorithm (on a UTM, or any other Turing-equivalent abstract machine) or *C*, could solve all problems in NP in polynomial time.

The following algorithm is used to remove noise i.e unwanted text from data. As the data file can be downloaded from internet, it may contain noise factors. This noise can be some unwanted punctuation marks, html tags and special characters.

**Algorithm: Noise Removal Algorithm**
**Input**: P here P is Document file
**Step 1**: d1 = Escaping htmlCcharacters (P)
**Step 2**: d2 = DecodingData (d1)
**Step 3**: d3= Apostrophe_Lookup(d2)
**Step 4**: d4= RemovalOfStopWords(d3)
**Step 5**: d5= Apostrophe_Lookup(d2)
**Step 6**: d6=Removal Punctuations(d5)
**Step 7**: d7=Removal Expressions(d6)
**Step 8**: d8= SplitAttachedWords(d7)
**Step 9**: d9= Slangslookup(d8)
**Step 10**: d10= StandardizingWords(d9)
**Step 11**: d11= RemovalOfUrl(d10)
**Step 12**: Stop

### *A. Hardware and Software Used*

Hardware Configuration

- Processor : -P-IV– 500 MHz to 3.0 GHz
- RAM : - 1GB
- Disk : -20 GB

Software Configuration

- Operating System: -Windows 7/XP
- Development End (Programming Languages):- Java

## V. EXPECTED RESULT

5.1 Dataset Information

Twelve multi-label data sets are used in the experiments. These data sets are benchmark data sets from different application domains: scene, emotions and corel5k are image data sets, genebase and yeast are biology data sets, and the remaining seven are document corpus. Reuters(10), Reuters(21), and Reuters(90) are the Reuters-21578 text data sets w.r.t. the largest 10 classes, 21 classes, and 90 classes. All the data sets are originally split into training and test set, and such originally given training/test data split are used in the experiments.

The practical implementation of proposed work and existing works is done using Java on real time public research datasets such as medical dataset.

As shown in figures 2, 3 and 4, the performance of accuracy is improved in proposed method, the processing time is decreased and also the recognition errors are minimized using proposed multi-label classification technique.

Table 1: Comparative Study of Existing & Proposed Methods

| ML Forest | IML Forest |
|---|---|
| It was mainly based on binary relevance which does not considered multi-label data. | Due to number of hierarchical trees are used multi-label data can be classified. |
| It does not consider noise factor present in dataset. | First noise is removed from dataset to obtain better results. |
| Margin of error is more due to presence of noisy data. | Margin of error is less since noise is removed from dataset. |

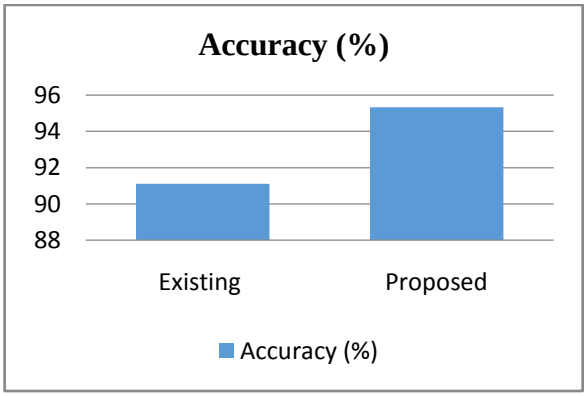


Figure 2: PerformanceAccuracy Evaluation

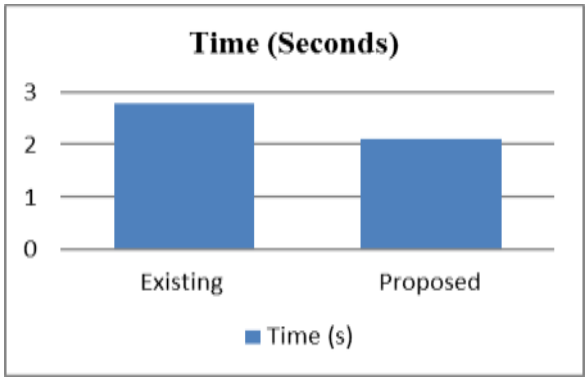


Figure 3: Performance Time Evaluation

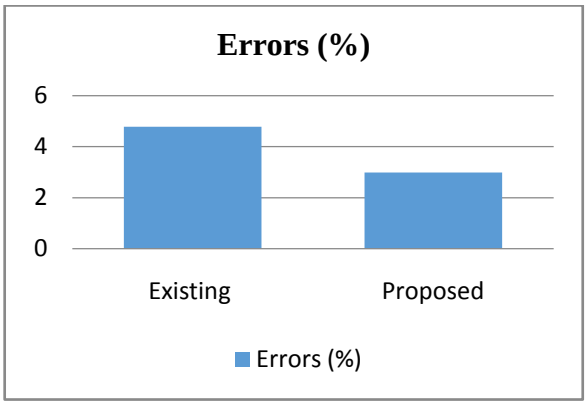


Figure 4: Performance Error Evaluation

## CONCLUSION AND FUTURE WORK

A new multi-label classification method with noise removal technique, called IML-FOREST is presented to build an ensemble classifier in this paper. In IML-FOREST, before constructing hierarchical trees noise is reduced from the dataset using noise removal algorithm, and a label transfer mechanism is developed which identifies the relevant labels hierarchically.A hierarchical multi-label classifier model can be very efficient on the tasks with a large number of labels if clustering technique is considered to organize the labels in growing the tree. This work remains to be implemented in future work.

## References

[1] T. N. Rubin, A. Chambers, P. Smyth, and M. Steyvers, "Statistical topic models for multi-label document classification," Machine learning, vol. 88, no. 1-2, pp. 157–208, 2012.

[2] J. Nam, J. Kim, E. L. Menc´ıa, I. Gurevych, and J. F¨urnkranz, "Large-scale multi-label text classification–revisiting neural networks," in Machine Learning and Knowledge Discovery in Databases, 2014, pp. 437–452.

[3] F. Sun, J. Tang, H. Li, G.-J. Qi, and T. S. Huang, "Multi-label image categorization with sparse factor representation, "Image Processing, IEEE Transactions on, vol. 23, no. 3, pp. 1028–1037, 2014.

[4] G. Tsoumakas and I. Katakis, "Multi-label classification: An overview," International Journal of Data Warehousing & Mining, vol. 3, no. 3, pp. 1–13, 2007.

[5] S. Huang, Y. Yu, and Z. Zhou, "Multi-label hypothesis reuse," in Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2012, pp. 525–533.

[6] K. Dembczynski, W. Waegeman, W. Cheng, and E. H¨ullermeier, "On label dependence in multi-label classification," in Workshop proceedings of learning from multi-label data, 2010, pp. 5–12.

[7] N. Cesa-Bianchi, C. Gentile, and L. Zaniboni, "Hierarchical classification: combining bayes with svm," in Proceedings of the 23rd international conference on Machine learning, 2006, pp. 177–184.

[8] ] M.-L. Zhang and K. Zhang, "Multi-label learning by exploiting label dependency," in Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining, 2010, pp. 999–1008.

[9] M. R. Boutell, J. Luo, X. Shen, and C. M. Brown, "Learning multilabel scene classification," Pattern recognition, vol. 37, no. 9, pp. 1757–1771, 2004.

[10] J. Read, A. Puurula, and A. Bifet, "Multi-label classification with meta-labels," in Data Mining (ICDM), 2014 IEEE International Conference on, 2014, pp. 941–946.

[11] Snehal D. Jadhav, Vaishali P. Suryawanshi, "A Review on Personalized Search Engine", in IJSWS- 15- 128, on 2015.

[12] YaminiKadwe, Vaishali P. Suryawanshi, "A Review On Concept Drift",IOSR Journal of Computer Engg, Volume 17, Issue 1, Ver 2 on Jan-Feb 2015.