I J C T A, 9(40), 2016, pp. 453-462 ISSN: 0974-5572

Hybrid Linear and RBF Kernel based Bio-NER System

Neeraj Battish* and Dapinder Kaur*

ABSTRACT

Biomedical Named Entity Recognition (BIO-NER) is a critical step of Text mining defined as an identification of biomedical text terms from huge amount of text data. Many technologies have been applied to this task in the past few years. However, Bio-NER remains challenging and still there is a huge gap between the best NER systems on biomedical literatures and the best algorithms in newswire domain. To deal with their problems, many approaches have been proposed, featured by three types: dictionary-based method, heuristic rule-based method, and statistical machine learning method. Currently, the statistical machine learning is most widely used as comparatively more robust and effective approach than others. The main process of the machine learning method is to construct proper models based on training data and use those models to classify the test data. Many models have been developed to identify biomedical entities, including hidden Markov model, support vector machine, maximum entropy, maximum entropy Markov model, and conditional random fields. So in this work a 'linear' and 'rbf" kernel based hybrid kernel for BIO-NER system has been proposed. This system is been implemented in MATLAB environment by using BIONER database. This system/model is a two phase model where firstly features is been extracted and selected by removing irrelevant terms like stop words. The second step of this enhanced model is classification where SVM Classifier is used with different kernels which helps to enhance the system and achieving 3-5% improved accuracy.

Keywords: BioNER, SVM, Kernel, Classification, Features;

1. INTROUCTION

With the increase in information in biomedical domain, there is a great demand for biomedical information extraction techniques. Recognising the entities such as RNAs, cells, DNAs etc. has become in biomedical knowledge discovery one of the important task. Though a lot of algorithms have been given for this purpose but NER [BIOMEDICAL NAMED ENTITY RECOGNITION] still remains a challenge and an area of active research [1], as still there is huge difference in F-score of 10 points between general newswire named entity recognition and biomedical named entity recognition. For biomedical NER it is more difficult in following ways:

- Biomedical NEs – most types do not have a complete dictionary and new NEs are being created continuously.
- Same phrase or word can point to different entities relying on their contexts. Biological NEs conversely have many spelling reform
- Quite often before NEs modifiers are used and biomedical NEs are sometimes very long. These points mark the difficulties for NEs boundary identification.
- NEs can be cascaded. Embedment of one NE can be done in another. For identification of these kinds of NEs more efforts must be made.

* Department of Computer Science and Engineering, CGC-College of Engineering, Landran, Mohali, India, *Emails: neerajbattish@yahoo.com, dapinder.coecse@cgc.edu.in*

In biomedical domain abbreviations are used quite often. As there are not many evidences in abbreviation for some NE class, it becomes difficult to classify them rightly to face these problems; it is required to explore rich features and effective methods. In biomedical literature there has been many trials to develop techniques to identify NE. They roughly categorise into three approaches- dictionary based approach, statistical machine learning based approach and heuristic rule based approach. However techniques for biomedical NER don't gain satisfactory results. Problems propose that individual biomedical NER system might not involve entity representations with a lot of rich features and no algorithm of single type is practical to gain best performance.

To judge output quality of NER system, many measures have been given. One possibility is accuracy on token level [2]. It is facing two problems, the huge majority of tokens in real world text are not included in entity names as generally defined, leading to baseline [always predict not an entity] accuracy which is extravagantly high more than 90%; and wrongly predicting full span of entity name is not correctly penalized [finding a person's first name when last name follows is marked as ½ accuracy].

A variant of F1 in academic conferences has been defined below:

- Firstly, Precision is the number of predicted entity name spans that line up exactly with spans in the gold standard evaluation data.
- Similarly Recall is the number of names in the gold standard that appear at exactly the same location in the predictions.
- F1 is combined meaning of above two.

It can be derived from above that any prediction has a wrong class if it misses a single token that is do not contribute to either recall or precision.

2. LEARNING & CLASSIFICATION

Along these lines information mining procedures, for example, grouping, affiliation and bunching are for the most part used to remove the covered up, beforehand inconspicuous learning from voluminous of databases. Of the different information examination method arrangement is a directed machine learning system which makes forecasts about future class cases by mapping cases of testing information to the predefined class marks which is gain from the supplied examples of classes with class names. There are a few models in characterizations, for example, probabilistic model, developmental algorithmic model and so forth. [3]

Characterization comprises of foreseeing a certain result in view of a given information. To anticipate the result, the calculation forms a preparation set containing an arrangement of traits and the separate result, typically called objective or forecast characteristic. The calculation tries to find connections between the traits that would make it conceivable to foresee the result. Next the calculation is given information situated not seen some time recently, called forecast set, which contains the same arrangement of properties, aside from the expectation property – not yet known. The calculation investigations the info and produces an expectation [5]. The expectation exactness characterizes how "great" the calculation is. Order procedure is equipped for preparing a more extensive mixed bag of information than relapse and is developing in fame.

Arrangement comprises of doling out a class name to an arrangement of unclassified cases.

1. *Administered Classification:* -In this, the arrangement of conceivable classes is known ahead of time.
2. *Unsupervised Classification:*-In this, set of conceivable classes are not known. After arrangement we can attempt to dole out a name to that class. Unsupervised order is called grouping.

Arrangement is an alternate method than grouping. Order is like grouping in that it likewise sections client records into unmistakable portions called classes. Yet, not at all like bunching, an arrangement examination obliges that the end-client/examiner know early how classes are characterized. It is vital that every record in the dataset used to manufacture the classifier as of now have a worth for the credit used to characterize classes [7]. Since every record has a worth for the credit used to characterize the classes, and on the grounds that the end-client settles on the ascribe to utilize, arrangement is a great deal less exploratory than grouping. The target of a classifier is not to investigate the information to find intriguing fragments, but instead to choose how new records ought to be characterized. Characterization schedules in information mining additionally utilize a mixture of algorithms.[11]

The information possessions of associations consistently stretch out to numerous terabytes of individual records that have collected over years of regularly impromptu or semi-arranged action. Organizations as often as possible don't have information arrangement systems and/or strategies set up that permit them to comprehend what information they hold and where it's found. This can have various pernicious impacts, for example,

- Finding documents can turn into a troublesome errand, which can affect an association's capacity to pick up the full advantages of the learning amassed in its information.
- Legal implications can come about because of not having the capacity to rapidly discover or produce archives for a court hearing.
- Security may be traded off if information isn't effectively coordinated to staff access profiles.
- You could be spending more cash than obliged if information isn't coordinated to capacity media of the proper expense per GB. Information characterization permits low-need information to be moved from elite stockpiling frameworks to optional media, for instance.

Throughout the years, new strategies from the machine learning field turned out to be more well known, deserting frameworks which utilize carefully assembled tenets [8]. Machine learning procedures permit the programmed affectation of tenet based frameworks or grouping naming calculations from distributed preparing information.

Throughout the years, new routines from the machine learning field turned out to be more mainstream, abandoning frameworks which utilize carefully assembled tenets. Machine learning strategies permit the programmed instigation of guideline based frameworks or arrangement naming calculations from assigned preparing information. This is accomplished by dissecting the discriminative elements of positive and negative samples. Comparative cases and redundancies happening in the information are converted into principles and henceforth pick up reflection over solid illustrations. Three distinct sorts of learning routines can be recognized by their necessities for the preparation information:

- Administered learning
- Semi-administered learning
- Unsupervised learning

Different classifiers do exist taking after are few of them:

- *SVM:* - SVM with Kernal capacity is a profoundly powerful model and functions admirably over an extensive variety of issue sets. It is a twofold classifier can be effectively reached out to multi-class order via preparing a gathering of parallel classifiers and utilizing "one versus all" or "one versus one" to anticipate. It is an intense procedure and perform best in an extensive variety of non-straight grouping issues [19]. Works extremely well when you have a little arrangement of information components on the grounds that it will grow the elements into higher measurement space, giving

that you additionally have a decent size of preparing information (generally, overfit can happen). SVM is not versatile in managing vast number (billions) of preparing information, so Logistic Regression with physically extended list of capabilities will be more down to business.

- *K Nearest Neighbor :-* is additionally called example based adapting yet not model-based learning in light of the fact that it is not realizing any model at all [13]. Training procedure is fundamentally retaining all the preparation information. For forecast of new information point, we discovered the nearest K neighbours from the preparation set and let them vote in favour of the last expectation. To focus the "closest neighbours", a separation capacity should be characterized (e.g. Euclidean separation is a typical one for numeric data variables). Voting can likewise be weighted among the K-neighbours in view of their separation from the new information point.
- *Neural Networks:* - Neural system is great at adapting non-direct capacity furthermore numerous yields can be learnt in the meantime. The preparation time is moderately long and it is likewise vulnerable to neighbourhood least traps [16]. This can be disposed of by doing numerous rounds and pick the best learned model

3. PROPOSED WORK

As a basic stride of content mining in biomedical writing, biomedical named element acknowledgment (Bio-NER) alludes to the distinguishing proof of biomedical content terms, including quality, protein, DNA, RNA, cell type, cell line and infection, etc. Just when biomedical named substances are accurately distinguished could other more unpredictable errands, for example, human quality/protein standardization and protein-protein connection extraction, be acknowledged viably. Numerous innovations have been connected to this assignment in the previous couple of years. Be that as it may, Bio-NER stays testing and there is still an expansive crevice between the best NER frameworks on biomedical writings and the best calculations in newswire area. To handle with their issues, numerous methodologies have been proposed, highlighted by three sorts: lexicon based technique, heuristic guideline based system, and factual machine learning strategy. At present, the measurable machine learning is most broadly utilized as a relatively more strong and viable methodology. The fundamental procedure of the machine learning system is to build legitimate models in light of preparing information and utilize those models to group the test information. Numerous models have been created to recognize biomedical substances, including shrouded Markov model, bolster vector machine, most extreme entropy, greatest entropy Markov model, and contingent arbitrary fields. The past work which took under thought presented a two stage model. Their two-stage technique isolates the errand into two subtasks: named element recognition (NED) and named element order (NEC). Examinations and correlations exhibit that their framework beats a large portion of the condition of-craftsmanship frameworks with a F-score of 76.06 percent. The principle center of proposed work is to enhance the exactness utilizing two-stage model with improved characterization process.

Different Phases of Proposed System: - This work is mainly focussed to identify the biomedical terms in the given sample. The main phases in this proposed system is given as:

Phase 1: Database Acquisition

Phase 2: Feature Extraction

Phase 3: Classification

Phase 4: Performance Analysis

The first step is data set selection that is known as data acquisition a data set is selected from Bio_ NER that means a bio medical named entity recognition system.In This Project GENIA Corpus is used for sample data. This data set obtained from http://www.nactem.ac.uk/tsujii/GENIA/ERtask/report.html.

In the next phase different algorithms for extracting different features form data set are implemented.

Features are as follows:

- Word
- Word shape
- Word length
- Suffix & Prefix
- Morphological

In the next phase classification process is performed. For implementing classification phase Enhanced SVM classifier is used which contains a hybrid kernel that is a combination f 'RBF' and 'Linear' kernel.

4. RESULTS & DISCUSSION

The very first step in this work is to acquire data. Like in this phase a data set is selected from a file named as sampletest1 from the system .This is the data to be trained and classified. This data set contains the test data set which contains the 5 entity classes: DNA, RNA, protein, cell_line and cell_type.

Feature Extraction: In this phase various algorithms are implemented for extracting different features form data set.

Features are: Word, Word shape, Word length, Suffix & Prefix, Morphological

Word: After selecting the data set the first algorithm introduced will work on **word level** all the joint words are retrieved from the corpus of words. These features has been saved in a cell form and then it can be extracted to store in the excel files.

Feature extracted and saved in excel file named as feature1.xls as shown above.

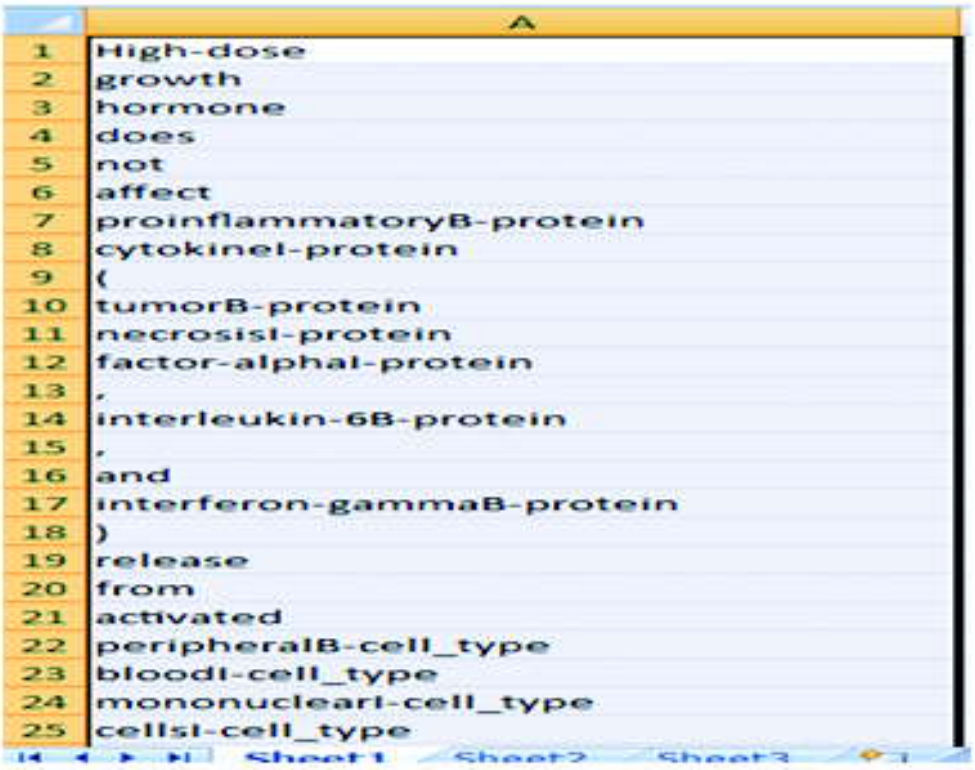


Figure 1: Excel sheet for word feature1.xls

Word shape: For this feature different rule has been followed to check the shape of the word. These rules are:

- Capitalized characters are replaced by "A"
- Capital Letters → A
- No capitalized characters are replaced by "a"
- No capitalized Letters → a
- Digits are replaced by "0"
- Digits → 0
- Non-English characters are unaltered.

In this step the algorithm will convert the word shapes into desirable shape here the requirement is to convert all the capitalized characters to 'A' and small to 'a'all numerals to '0'rest of the digits stay as it is.

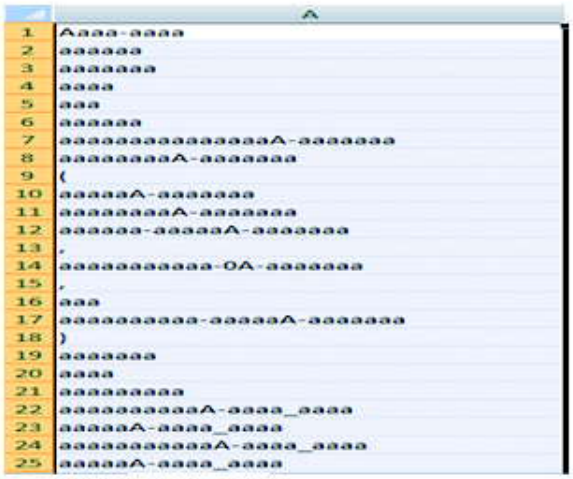


Figure 2: Excel sheet for word shape feature2.xls

Feature extracted and saved in excel file named as feature2.xls as shown above.

Suffix & Prefix

For each word, the two &three characters' prefix and suffix are used as features. There is given a prefix count it could be 2 or 3.The user has given 2 so from each word extracted earlier a prefix of two characters will be extracted further to perform this step. The prefix characters will be stored in the excel file shown below

Figure 3: Excel sheet for prefix words

The working of 2 characterSufixis similar to prefix step.

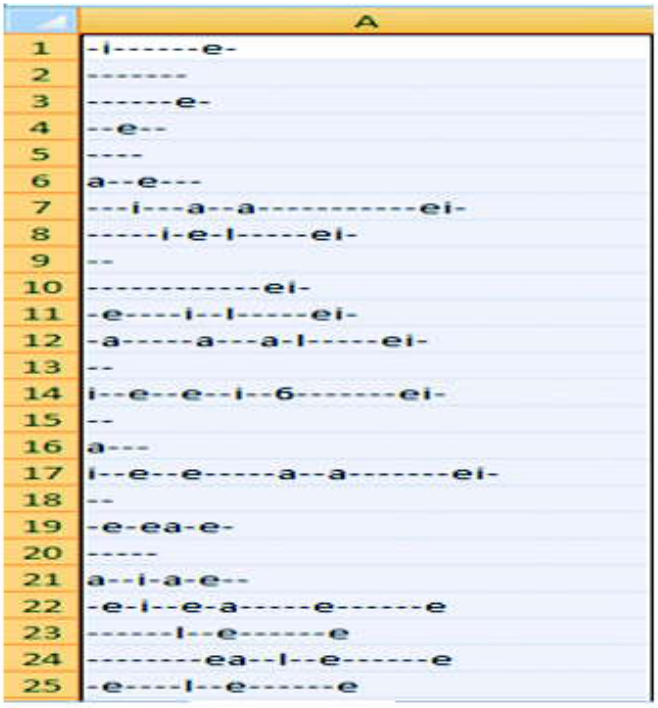
	A
1	e
2	he
3	es
4	st
5	tr
6	in
7	in
8	(in
9	inn
10	in
11	,in
12	,d
13	in
14	)e
15	m
16	d
17	pe

Figure 4: Excel sheet for suffix

Morphological Feature II

For this feature each character of the current word is replaced by "-," except five vowels (i.e., a,e, i, o, u).

For Example, DNA → —A. This is called as morphological features only the vowels are treated as special characters and are not changed to "-", Rest of the characters are changed to "-". Fig. 6 shows the converted data sample pass to the system.



	A
1	-i------e-
2	-------
3	------e-
4	--e--
5	----
6	a--e---
7	---i---a--a----------ei-
8	-----i-e-l-----ei-
9	--
10	----------ei-
11	-e----i--l-----ei-
12	-a-----a---a-l-----ei-
13	--
14	i--e--e--i--6-------ei-
15	--
16	a---
17	i--e--e-----a--a-------ei-
18	--
19	-e-ea-e-
20	-----
21	a--i-a-e--
22	-e-i--e-a-----e------e
23	------l--e------e
24	--------ea--l--e------e
25	-e----l--e------e

Figure 5: Excel sheet for morphological feature extraction

Classification

As we have seen, the undertaking of NERC has grown throughout the years and moreover has the connected routines. One noteworthy objective of the grouping is to make preparing information accessible to machine learning frameworks. Named elements are obscure words on the grounds that they can't be turned upward in any customary vocabulary. To distinguish them in a machine learning situation, an arrangement of particular components is expected to tell positive and negative illustrations separated.

Throughout the years, new routines from the machine learning field turned out to be more mainstream, deserting frameworks which utilize high quality tenets. Machine learning procedures permit the programmed actuation of principle based frameworks or grouping naming calculations from allotted preparing data [25]. This is accomplished by breaking down the discriminative components of positive and negative cases. Comparative cases and redundancies happening in the information are converted into tenets and consequently pick up reflection over solid cases. Three distinct sorts of learning strategies can be recognized by their necessities for the preparation information:

- Supervised learning
- Semi-supervised learning
- Unsupervised learning

For Recognition SVM Classifier is utilized as a part of this work. Support vector machines are computational calculations that develop a hyperplane or an arrangement of hyperplanes in a high or unbounded dimensional space. SVMs can be utilized for order, relapse, or different assignments. Naturally, a division between two straightly detachable classes is accomplished by any hyperplane that gives no misclassification on all information purposes of any of the considered classes, that is, all focuses having a place with class An are named as +1, for instance, and all focuses having a place with class B are marked as - 1.

This methodology is called linear classification however there are numerous hyperplanes that may characterize the same arrangement of information.

Portion capacity svm train uses to delineate preparing information into piece space. The default bit capacity is the dab item. The part capacity can be one of the accompanying strings or a capacity handle:

- 'linear' — Linear bit, importance speck item.
- 'quadratic' — Quadratic bit.
- 'polynomial' — Polynomial bit (default arrange 3). Indicate another request with the polyorder name-worth pair.
- 'rbf' — Gaussian Radial Basis Function portion with a default scaling variable, sigma, of 1. Indicate another worth for sigma with the rbf_sigmana.

RBF kernel in SVM

In machine taking in, the (Gaussian) outspread premise capacity piece, or RBF part, is a famous bit capacity utilized as a part of different kernel zed learning algorithms [25]. Specifically, it is generally utilized as a part of bolster vector machine characterization. The functions are nonlinear. In a pattern recognition task this often guarantees better separation. The functions map the samples from the input space to hidden space, in which the same samples are hopefully separable.

Table 1
Comparison table of three techniques

Technique	F-score
SVM [linear kernel]	76%
SVM [rbf kernel]	79%
E-SVM [hybrid kernel]	84%

This is the tabular representation of the difference between the F-score values of Basic Technique, SVM and E-SVM.As it is very much clear from the table given above that E-SVM is providing the maximum value for F-score 84% and for efficient classification F-score value must be higher so E-SVM proved to be a good way to achieve better classification results.

5. CONCLUSION

This work is focused on performance of classifier model for Biomedical Named Entity Recognition. In this work SVM is used for quick identification of biomedical entity. This SVM classifier has been used where training function uses two kernels named as linear and rbf kernel and this is compared with rbf kernel only. The data set of GENIA Corpus used for this work which includes different biomedical entities like protein, cell_type, DNA, RNA and Cell_line. This is two phase system where named entity detection process includes pre-processing and feature extraction and other is named entity recognition where classification takes place. The above results conclude that enhanced classifier works better as it gives 84% F-Score whereas existed gives 76% and 79% respectively. These results verify the accuracy of the enhanced classifier. Future work could go in direction to test this classifier with different data sets so that their accuracy can be measured.

REFERENCES

[1] M. Dalla Mura, A. Villa, J. A. Benediktsson, J.Chanussot, and L. Bruzzone, ¯Classification of hyperspectral images by using extended morphologicalattribute profiles and independent component analysis,"IEEEGeosci. Remote Sens. Lett., Vol. 8, pp.no 542–546,May 2011.

[2] Maria Vargas Vera "Knowledge Extraction by using an Ontology based Annotation Tool" Knowledge Media Institute (KMi), The Open University,Walton Hall, Milton Keynes, MK7 6AA, United Kingdom .

[3] BalajiPadmanabhan,"Data Mining Overview and Optimization Opportunities".Microsoft Research Report MSR-TR-98-04, January 2013.

[4]] Pang Ning Tan, M. Steinbach and V. Kumar, "Introduction to Data Mining" ,Addison Wesley, pp. no.769,June 2005.

[5] Chen, K. and Liu. L. A random rotation perturbation approach to privacy data classification. In Proc of IEEE Intl. Conf. on Data Mining (ICDM), pp.no.589-592,December 2005.

[6] Y. Li, C. Zhang, and J.R. Swan, "An Information Filtering Model on the Web and Its Application in Jobagent," Knowledge-Based Systems, vol. 13, no. 5, pp. 285-296, 2000.

[7] Geniahttp://www.nactem.ac.uk/genia/

[8] Yu Qiao,HuiPingLiua, MuiBai, XiaoDong Wang, XiaoLuo Zhou, "The decision tree algorithm of urban extraction from multisourceimage data".Vol. 3, pp. 301-308, June 2005.

[9] Krishnalal G, S BabuRengarajan and K G Srinivasagan "A New Text Mining Approach Based on HMM-SVM for Web News Classification" International Journal of Computer Applications (0975 - 8887) Volume 1, No. 19, pp. 98-104, 2010.

[10] Kalyani Manda1, Suresh Chandra Satapathy2, B. Poornasatyanarayana,International Journal Electronics Signals and sytems Vol. 2,Issue 7,ISSN:2249-9482,July 2012.

[11] Lishuang Li, Wenting Fan, and Degen Huang "A Two-Phase Bio-NER System Based on Integrated Classifiers and Multiagent Strategy" IEEE/ACM Transactions on Computational Biology and Bioinformatics, Vol.10, No.4, pp. 897-904, 2013.

[12] Jongwoo Kim*, Daniel X. Le, George R. Thoma "Identification of Investigator Name Zones using SVM Classifiers and Heuristic Rules" 12th International Conference onDocument Analysis and Recognition (ICDAR), 2013, pp. 140-144.

[13] Zhihua Liao "Biomedical Named Entity Recognition Based on Skip-Chain CRFS" International Conference onIndustrial Control and Electronics Engineering (ICICEE), 2012, pp. 1495-1498.

[14] B V Chowdary, Annapurna Gummadi, UNPG Raju,BAnuradha and RavindraChangala "Decision Tree Induction Approach for Data Classification Using Peano Count Trees" International Journal of Advanced Research in Computer Science and Software Engineering, Vol. 2, No. 4, pp. 475-479, April 2012

[15] C. M. Velu, Kishana R. Kashwan "Performance Analysis for Visual Data Mining Classification Techniques of Decision Tree, Ensemble and SOM" International Journal of Computer Applications (0975 – 8887), Volume 5, No.22, pp. 65-71, November 2012.

[16] S .subbaiah "Extracting Knowledge using Probabilistic Classifier for Text Mining" International Conference on Pattern Recognition, Informatics and Mobile Engineering, February 21-22 2013.

[17] C NamrataMahender "Text Classification and Classifiers a Survey" International Journal of Artificial Intelligence & Applications (IJAIA), Vol.3, No.2, March 2012

[18] Annetietenteije "Knowledge Engineering and Knowledge Management" 18th International Conference, EKAW 2012, Galway City, Ireland, October 8-12, 2012

[19] TipawanSilwattananusarn "Data Mining and Its Applications for Knowledge Management:A Literature Review from 2007 to 2012" International Journal of Data Mining & Knowledge Management Process (IJDKP) Vol.2, No.5, September 2012.

[20] Raymond J. Mooney "Mining Knowledge from Text Using Information Extraction" Department of Computer Sciences University of Texas at Austin 1 University Station C0500 Austin, TX 787120233 Volume 7, Issue 1, pp 10, 2010.

[21] Suneetha K.R "Data Preprocessing and Easy Access Retrieval of Data through Data Ware House" Proceedings of the World Congress on Engineering and Computer Science, 2009, Vol I, October 20-22, 2009, San Francisco, USA.

[22] Li Fu1+, Jian Cheng1, 2, 3 and Yongheng Zheng1,"Object-oriented Classification of High-resolution Remotely Sensed Imagery" IPCSIT Vol. 47,pp. no.123,july 2010.

[23] MrinaliniRana.,"A survey of association rule mining using genetic algorithm", International journal of computer application & information Technology Vol.1, IssueII, ISSN:2278, pp. no.1-8,Auguest 2012.

[24] Mónica Marrero, JuliánUrbano, Sonia Sánchez-Cuadrado, Jorge Morato, Juan Miguel Gómez-Berbís "Named Entity Recognition: Fallacies, Challenges and Opportunities" Computer Standards & Interfaces Volume 35, Issue 5, September 2013, pp. 482–489.

[25] Nadeau, David, and Satoshi Sekine. "A survey of named entity recognition and classification."LingvisticaeInvestigationes 30, no. 1, pp. 1-20, (2007).

[26] P.Pooja, J.Jayanthv and S. Koliwad, "Classification of RS data using Decision Tree Approach," International Journal of Computer Applications,Vol. 23(3), pp. no.7-11, Febuary2011