Performance of machine learning methods for classification tasks

B. Krithika
M.E., CSE, Final Year Student,
Dept. of CSE, Annamalai University,
Annamalainagar.
email id:krithima13@gmail.com

Dr. V. Ramalingam
Professor and Head,
Dept. of CSE, Annamalai,
University, Annamalainagar,
email id: aucsevr@gmail.com

K. Rajan
HOD, Dept. of CSE,
Muthiah Polytechnic College,
Annamalainagar,
email id: kaliyaperumalrajan@yahoo.co.in

Abstract

In this paper, the performance of various machine learning methods on pattern classification and recognition tasks are proposed. The proposed method for evaluating performance will be based on the feature representation, feature selection and setting model parameters.

The nature of the data, the methods of feature extraction and feature representation are discussed. The results of the Machine Learning algorithms on the classification task are analysed. The performance of Machine Learning methods on classifying Tamil word patterns, i.e., classification of noun and verbs are analysed.

The software WEKA (data mining tool) is used for evaluating the performance. WEKA has several machine learning algorithms like Bayes, Trees, Lazy, Rule based classifiers.

KEYWORDS : Machine learning, pattern classification, pattern recognition, feature representation, feature selection, setting model parameters, Tamil word patterns, noun, verbs and Weka.

I. INTRODUCTION

A. Introduction to Machine Learning

Machine learning is the subfield of artificial intelligence that is concerned with the design and development of algorithms that allow computers to improve their performance over time based on data, such as from sensor data or databases. Machine learning is closely related to fields such as data mining, statistics, inductive reasoning, pattern recognition, and theoretical computer science.

Machine learning is programming computers to optimize a performance criterion using example data or past experience.

Machine learning algorithms are organized based on the desired outcome of the algorithm. Common algorithm types include:

- Supervised Learning
- Un-supervised Learning
- Semi- supervised Learning
- Reinforcement learning
- Transduction
- Learning to learn

In this paper, the performances of various Machine learning techniques available in WEKA are discussed.

B. Tamil Language

Tamil grammar is agglutinative in nature. Suffixes are used to mark class, number and cases attached to a noun. Tamil word may have a lexical root to which one or more affixes are attached. Most of the Tamil affixes are suffixes which can be derivational or inflectional. Length and extent of agglutination is longer in Tamil resulting in longer words with many suffixes. Some of the other issues are morpho-phonology rules, complex noun and verb patterns, and out of vocabulary rate due to inflections. Poetry forms are more complex than prose forms.

In Tamil, nouns are classified into rational and irrational forms. Humans come under rational form whereas all other nouns are classified as irrational. Rational nouns and pronouns belong to one of the three classes: masculine singular, feminine singular and rational plural. Irrational nouns belong to one of the two classes: irrational singular and irrational plural. Suffixes are used to perform functions of cases or post positions. Tamil verbs are also inflected through the use of suffixes. The suffix of the verb indicates person, number, mood, tense and voice.[Selvam, Natarajan,[12]].

Tamil is consistently head-final language. The verb comes at the end of the clause with a typical word order of Subject Object Verb (SOV). However, Tamil allows word order to be changed making it a relatively word order free language. Other features are plural for honorific noun, frequent echo words, and null subject feature i.e. all sentences do not have subject, verb and object.

C. Pattern Classification

Pattern classification is the organization of patterns into groups of patterns sharing the same set of properties.

Automatic (machine) recognition, description, classification, and grouping of patterns are important problems in a variety of engineering and scientific disciplines such as biology, psychology, medicine, marketing, computer vision, artificial intelligence, and remote sensing.

The design of a pattern recognition system essentially involves the following three aspects:

i) data acquisition and preprocessing,

ii) data representation, and

iii) decision making.

D. Statistical Pattern Recognition

A pattern is represented by a set of ***d*** features, or attributes, viewed as a ***d***-dimensional feature vector.

The recognition system is operated in two modes: training (learning) and classification (testing).

In the training mode, the feature extraction/selection module finds the appropriate features for representing the input patterns and the classifier is trained to partition the feature space.

In the classification mode, the trained classifier assigns the input pattern to one of the pattern classes under consideration based on the measured features. [Anil K. jain [1]]

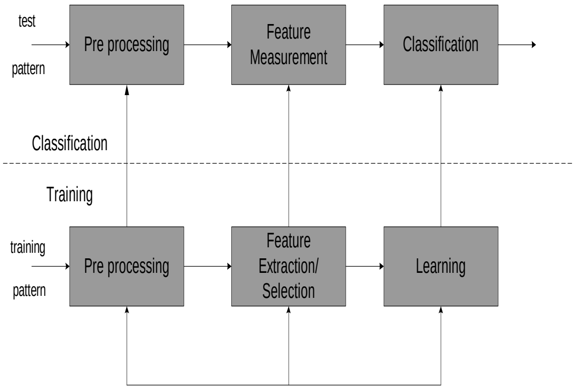

Figure 1 Model for Statistical Pattern Recognition

E. Syntactic Classification

A syntactic category is a set of words and/or phrases in a language which share a significant number of common characteristics. The classification is based on similar structure and sameness of distribution (the structural relationships between these elements and other items in a larger grammatical structure), and not on meaning.

F. Objectives of this Paper

- To classify the Tamil words into verb and noun.
- To extract the features from the Tamil words.

- The extracted features are given to the software we used.
- To tabulate the performances of machine learning algorithms for noun and verb classification.

G. Outline of this Paper

This paper is organized as follows. In Section II proposes related works. Section III describes methodology. Section IV presents the experimental results and discussion. Section V concludes this paper.

II. RELATED WORKS

Part of speech tagging (POS) is the task of labeling each word in a sentence with its appropriate syntactic category called Part of speech. POS tagging is a very important pre-processing task for language processing activities.

POS taggers for Indian languages like Malayalam, Bengali, telugu, Punjabi, and hindi were reported.

A stochastic Hidden Markov Model and Support Vector Macine based part of speech tagger is used for Malayalam [Manju K., Soumya S., Suman Mary Idicula [3]].

In case of Bengali Language three taggers have been proposed. All the proposed taggers used different tagging approaches for doing POS tagging. Hidden Markov Model and Maximum Entropy (ME) based stochastic taggers were proposed [Sandipan Dandapat, Sudeshna Sarkar, Anupam Basu [4]]. Support Vector Machine was also proposed [Ekbal, A. Bandyopadhyay, S., [5]].

In case of Hindi language different POS tagging approaches have been proposed [Aniket Dalal, Kumar Nagaraj, Uma Sawant And Sandeep Shelke [6]]; [Smriti Singh, et.al, [7]]. Morphology driven tagger [Smriti Singh, et.al, [7]], Maximum Entropy based tagger [Aniket Dalal, Kumar Nagaraj, Uma Sawant And Sandeep Shelke [6]], HMM based tagger [Manish Shrivastava and Pushpak Battacharyya [8]] and Conditional Random Field based tagger [John Lafferty, Andrew McCallum, and Fernando Pereira [11]] have been proposed for Hindi language.

In case of Punjabi language a rule based part-of-speech tagging approach was used, which is further used in grammar checking system [Singh Mandeep, Lehal Grupreet, and Sharma Shiv [10]].

In case of Telugu language, three POS taggers have been proposed Rule-based approach, using transformation based learning (TBL) approach of Erich Brill and using Maximum Entropy model, a machine learning technique [RamaSree, R.J, Kusuma Kumari, P., [9]].

III. METHODOLOGY

In this paper, Tamil words are classified. The majority of the words in any language are nouns and verbs. These words are inflected by adding more number of suffixes. So, identification of each word is very difficult. The dictionaries/lexicons cannot have all possible word forms of the languages. The dictionaries usually have listed only the root words.

The word classification begins with the dictionary look up. After finding the longest stem of the given word, the stem and suffixes are separated.

The major category of the word is the category of the stem. The subcategory of the word is determined by checking the suffixes.

A. Verb

Verbs can be subdivided into different types based on morphological and syntactical point of view. Verbs are primarily classified into finite verb and non-finite verb.

The verbs are classified into different classes based on the type of first suffix they take. Tamil verb can be divided into several classes. A number of classifications have been suggested in the literature. The classification is shown below.

Table 1 Verb Classes

Class	Present	Past	Future
I (செய்)	கிற்	த்	வ்
II (உட்கார்)	கிற்	ந்த்	வ்
III (தூங்கு)	கிற்	இன்	வ்
IV (போடு)	கிற்	Doubling	வ்
V (நில்)	கிற்	ன்ற்	ப்
VI (படி)	க்கிற்	த்த்	ப்ப்
VII (நட)	க்கிற்	ந்த்	ப்ப்

B. NOUN

A noun is a part of speech typically denoting a person thing, place or idea.

Table 2 Noun Paradigm

Case	Singular	Plural
Nominative	புத்தகம்	புத்தகங்கள்
Accusative	புத்தகத்தை	புத்தகங்களை
Instrumental	கருவியால்	கருவிகளால்
Dative	அவனுக்காக	அவர்களுக்காக
Ablative	புத்தகத்திலிருந்து	புத்தகங்களிலிருந்து
Genitive	புத்தகத்தின்	புத்தகங்களின்
Locative	புத்தகத்தில	புத்தகங்களில்
Sociative	புத்தகத்துடன	புத்தகங்களுடன்

C. Feature Extraction

The machine learning algorithms require input file which contains features and class labels. For our word classification problem we use the following features for each inflected noun.

i) Categories of root: Hn, Nhn, Nmn, An, Ian, Abn

ii) The characters following the root.

A maximum of 15 characters following the root are used.

Empty character (x) is used for shorter words.

iii) Class labels : plu, acu,emp, gen

eg., மரத்தை – மர(ian) த் த் ஐ x x x x x x

x x x x x x→ acu

D. Features for Noun Classification

cat1	c1	c2	c3	c4	c5	c6	c7	c8	c9	c10	c11	c12	c13	c14	c15	cat2
abn	த்	த்	இ	ன்	X	X	X	X	X	X	X	X	X	X	X	gen
ian	த்	த்	ஐ	X	X	X	X	X	X	X	X	X	X	X	X	acc
abn	ஆ	க்	அ	X	X	X	X	X	X	X	X	X	X	X	X	pur
abn	ஆ	ய்	X	X	X	X	X	X	X	X	X	X	X	X	X	adv
abn	ஆ	ன்	அ	X	X	X	X	X	X	X	X	X	X	X	X	adj
abn	ஆ	ன்	அ	த்	உ	X	X	X	X	X	X	X	X	X	X	par
abn	ஆ	ன்	ஆ	ல்	X	X	X	X	X	X	X	X	X	X	X	comp
ppn	ம்	ஏ	X	X	X	X	X	X	X	X	X	X	X	X	X	emp
ian	ங்	க்	அ	ள்	X	X	X	X	X	X	X	X	X	X	X	plu
ian	ஆ	ய்	X	X	X	X	X	X	X	X	X	X	X	X	X	adv
ian	ஆ	ன்	அ	X	X	X	X	X	X	X	X	X	X	X	X	adj
abn	ங்	க்	அ	ள்	X	X	X	X	X	X	X	X	X	X	X	plu
abn	த்	த்	இ	ல்	X	X	X	X	X	X	X	X	X	X	X	loc
abn	த்	த்	இ	ற்	க்	உ	X	X	X	X	X	X	X	X	X	dat
ian	உ	க்	க்	உ	ம்	X	X	X	X	X	X	X	X	X	X	dat
ian	உ	ட்	அ	ன்	X	X	X	X	X	X	X	X	X	X	X	soc
abn	க்	அ	ள்	X	X	X	X	X	X	X	X	X	X	X	X	plu

Output class labels: Gen-genitive, acc-accusative, adv-adverb, adj-adjective, emp-emphatic, plu-plural, loc-locative, dat-dative, abl- ablative, soc- sociative.

The category of the stem is labeled as ***cat1*** and the remaining 15 symbols are labeled as (*c1*, *c2*...*c15*). Based on the length of the word the symbols are either a Tamil character or an empty symbol(X). The output is labeled as ***cat2.***

These instances are made as training and testing instances. These instances are further given to data mining tool we used. The performance of various learning algorithms is discussed.

E. Features for Verb Classification

Segment the stem (verb) from the word and segment the continuous suffixes according to the length of the word. We are considering 3 characters or symbols from *c1, c2 and c3.* Stem(verb) and characters from *c1,c2and c3* is collectively called as featured vector. These 3 features are given as input and type is taken as output. In total, there are 4 features which are given as input data for training. Output type is present and past tense suffixes. Present tense suffixes are è¤ø¢ and è¢è¤ø¢.

F. Feature Representation

Most of the machine learning algorithms accepts nominal data as features. So the characters are given directly as a feature. The characters can be represented by unique number (usually ASCII) for algorithms which require numerical data.

IV. RESULTS AND DISCUSSION

The performance of the classifiers are based on correctly and in correctly classified instances, kappa statistic, Mean absolute error, Root mean squared error, Relative absolute error and Root relative squared error.

The general set up of our experiments is the following. Each experiment is done using a 10-fold cross-validation on the available data. This means that the data is split in 10 partitions, and each of these is used once as test set, with the other nine as corresponding train set. We use default settings.

Table 3 Accuracy of Different Classifiers (For Noun Classification)

Model	Accuracy (%)
BAYES	90.60
LAZY	94.00
TREES	94.23
RULES	89.23

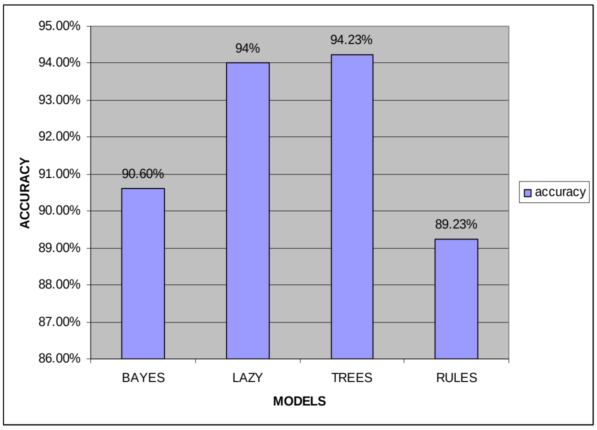

Figure 2 Accuracy of groups of different classifiers for Noun Classification

Table 4 Accuracy of Different Classifiers (For Verb Classification)

Model	Accuracy (%)
BAYES	86.00
LAZY	87.44
TREES	87.82
RULES	86.00

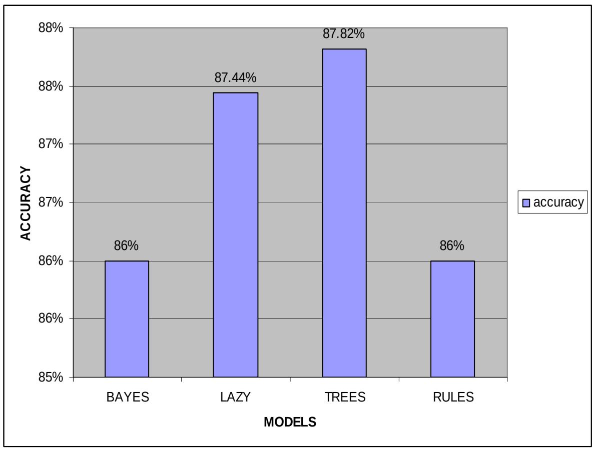

Figure 3 Accuracy of groups of different classifiers for Verb Classification

V. CONCLUSION

In this paper, the performance of various machine learning algorithms on classification of Tamil words was studied.

We used nouns and verbs from Tamil language text. We discussed different feature extraction and representation methods. Features are extracted from the stem (root) and suffixes of the words given. The extracted character level features are represented as nominal data as well as numerical data. These two types of feature representation schemes were used for preparing the training and test data.

In this paper, Bayes, Trees, Rule based classifiers and Lazy types of classifiers are studied. Each classifier has different learning algorithms. The performance of various algorithms is tabulated.

On the given set of features, we observed that the performances of Tree classifiers are better than other types of classifiers on both noun and verb. The performance obtained on noun and verb classifications are 94.23 and 87.82 respectively.

REFERENCES

[1] Anil K. Jain, Robert P.W. Duin, and Jianchang Mao, “Statistical Pattern Recognition: A Review”, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 22, no. 1, January 2000.

[2] Walter Daelemans, “Evaluation of Machine Learning Methods for Natural Language Processing Tasks” In Proceedings of the Third International Conference on Language Resources and Evaluation (LREC 2002). Las

[3] Manju K., Soumya S., Suman Mary Idicula, “Development of a POS tagger for Malayalam – An Experience,” artcom, International conference on advances in recent technologies in communication and computing, 2009, pp. 709-713.

[4] Sandipan Dandapat, Sudeshna Sarkar, Anupam Basu, “Automatic Part-Of-Speech Tagging for Bengali: An Approach for Morphologically Rich Languages in a Poor Resource Scenario”, Proceedings of the Association for Computational Linguistic, 2007, pp 221-224.

[5] Ekbal, Asif and Sivaji Bandyopadhyay, “Part of Speech Tagging in Bengali using Support Vector Machine”, ICIT-08, IEEE International Conference on Information Technology, 2008 pp.106-111.

[6] Aniket Dalal, Kumar Nagaraj, Uma Sawant and Sandeep Shelke, “Hindi Part-of-Speech Tagging and Chunking: A Maximum Entropy Approach”, In Proceeding of the NLPAI Machine Learning Competition, 2006.

[7] Smriti Singh, et.al, “Morphological Richness Offsets Resource Demand-Experiences in Constructing a POS Tagger for Hindi”, In the Proceedings of COLING/ACL, 2006, pp. 779-786.

[8] Manish Shrivastava and Puspak Bhattarcharyya, Hindi POS Tagger Using Naïve Stemming: Harnessing Morphological Information without Extensive Linguistic Knowledge, International Conference on NLP (ICON 08), Pune, India, December, 2008. Also accessible from http://ltrc.iiit.ac.in/proceedings/ICON-2008.

[9] RamaSree, R.J. Kusuma Kumari, P., “Combining Pos Taggers For Improved Accuracy To Create Telugu Annotated Texts For Information Retrieval”, 2007. Available http://www.ulib.org/conference/2007/Rama sree.pdf.

[10] Singh Mandeep, Lehal Gurpreet, and Sharma Shiv, “A Part-of-Speech Tagset for Grammar Checking of Punjabi”, Published in the Linguistic Journal, Vol. 4, no. 1, pp. 6-22, 2008.

[11] Lafferty, J., McCallum, A., and Pereira F., “Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data”, In Proc. of the 18th ICML’01, 2001, pp. 282-289.

[12] M. Selvam, A.M. Natarajan, “Improvement of Rule Based Morphological Analysis and POS Tagging in Tamil Language via Projection and Induction Techniques”. INTERNATIONAL JOURNAL OF COMPUTERS, Vol. 3, no. 4 2009.