A REVIEW OF TAMIL EPIGRAPHY DATA ANALYSIS THROUGH MACHINE LEARNING

S. Shahul Hammed [1], Dr. B. Arun Kumar[2]

ABSTRACT

Machine learning is an important and growing area in the field of Artificial intelligence. in which the machine identifies the next task to perform by analyzing the sample data given. Tamil epigraphy and corpora contain huge amount of data. The data contain varieties of Tamil language and different grammatical structures of the different generations of Tamil speakers. Previous methods

of analyzing the synchronic study of Tamil language and the diachronic comparison of words in texts used rule based POS (Parts of Speech) tags Database mapping. In this paper, I review and compare various machine learning methods to improve the corpus analysis to identify different varieties of Tamil epigraphy and understand the grammatical structure of each generation of Tamil.

Keyword : Machine learning, Epigraphy, Data mining, POS,

I. INTRODUCTION

Machine learning is a different method to analyze the problems in the area of artificial intelligence. It is a method of teaching and improves the knowledge of machines,-; based on the knowledge the machines make the decisions effectively and efficiently. Machine learning attempts to find the prediction automatically based on past experiences. [1].

The objective of machine learning is to learn from the data and information. Many studies have shown how machines are learning. Using some data to write the program to identify the solution for the problem is not machine learning, it is called automation. Machine learning and automation are different. For Making computers to think and make decisions like human beings, the decisions should be made on the basis of stored knowledge rather than mechanical solutions. [2]

Tamil epigraphy and corpus analysis contain a vast database. The database contains different varieties of Part of Speech (POS) tags. In Tamil epigraphy and corpus analysis, some methods are using mapping of POS tags database to identify the relationship between different varieties of Tamil and Mapping the grammatical structure of each generation of Tamil varieties. The method uses mapping and automation. This method does not make efficient and fine-grained analysis of Tamil epigraphy.

Another problem in this method is the difficulty in forming greater number of POS tags [3] [4]. With the availability of abundance of datasets in Tamil

[1]Assistant professor, Department of Computer Science and Engineering, Karpagam Academy of Higher Education, Coimbatore
Email:shahul.y2s@gmail.com

[2]Associate professor, Department of Computer Science and Engineering, Karpagam Academy of Higher Education, Coimbatore
Email:arunkumar.oct06@gmail.com

epigraphy, the demand of machine learning is rising. Many machine learning methods are available to find good solutions for this problem. The aim of this paper is to identify a good machine learning method for Tamil epigraphy and corpus analysis.

Machine learning methods are explained in section II, details about the suitable method for Tamil epigraphy and corpus analysis in section III, and conclusion are given in Section IV.

II. METHODS OF MACHINE LEARNING

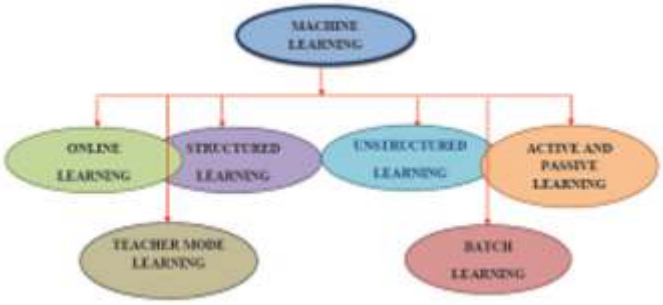

Fig-1-Machine Learning methods

1. STRUCTURED LEARNING :

The method uses a rule set for mapping the input and output parameters to find decisions and keep monitoring the rule set[5]. For example to buy a drumstick in the market, if it is soft and greenish, we buy it. If its hard and brown we don't. Deciding factors for buying a drumstick is color and softness. This is called domain set(A), the decision (buy or not) is called labels(B). The mapping function links the labels(B) and domain set (A) is called rule set(A->B). If we teach these rules set to the learner (machine), based on the rule set, it makes decisions for new inputs is called predictor. Based on the rule set the learner makes decisions using classification and regression. In this method applying the rule based(labels) data analysis of Tamil Epigraphy grammar gives some results depending on the input structure of the data, and identifies the variety of the Tamil language structure.

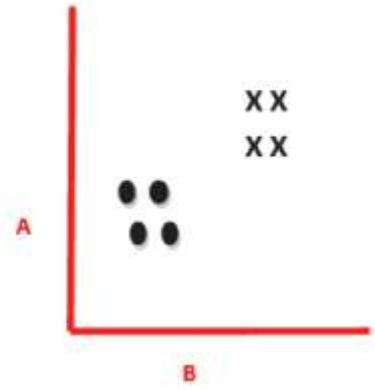

Fig-2 (Structured learning)

2. UNSTRUCTURED LEARNING :

An unstructured learning method examines only the values of Input. There is no rule for the value of the output. From the input it finds the pattern and will expose us to the output. There are two classifications in unstructured learning, namely clustering and association. The method of analyzing and Categorizing similar data is called clustering, for example the sale of a product per day in a super market. Using the sale information (data) to analyze how much of that product is sold per day, as input it is easy to identify how much will be needed the following day. Taking the details that have been discovered from the clustering method to apply on the similar products is called association. By using similar options, we can increase the sale of various products. The combination of structured and unstructured learning is called semi-structured learning. Tamil Epigraphy contains a vast amount of data. It is impossible to construct rule (label) for all data. Many important data should be labeled using structured learning and the pattern for remaining data should be found out using unstructured learning. The Semi-structured learning is more helpful to

analyze the Tamil epigraphy data using rule and pattern method.

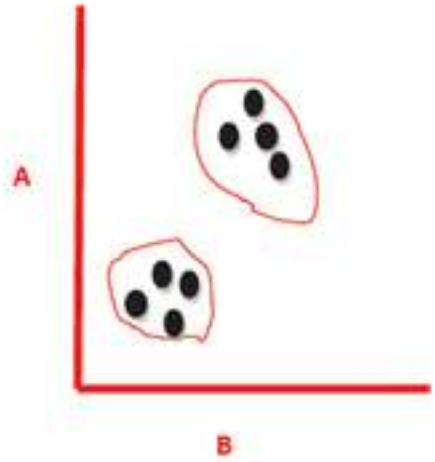

Fig - 3 (Unstructured Learning)

3. ACTIVE AND PASSIVE LEARNING :

Exploring and learning according to the given rules depend on the input data which is called passive learning. Example an email identifies the spam mail using the rule. The received mail it can't be confirmed as a normal email or spam mail., In these situations, It puts question to the user to clear the doubt and adding that as a new rule is called active learning[6].

4. TEACHER MODE LEARNING :

When the rules are violated, the arrangement of someone acting like a teacher (Monitor) to show which is the correct method(rule) to process the data is called teacher mode learning. The examples are Spam filtering, malware Detection and biometric recognition.

5. ONLINE AND BATCH LEARNING :

In this method data come in a linear order and are used to predict and update future data at every step, in contrast to batch learning method [7]., Online learning is a general method used in areas of machine learning, where it is computationally not possible to train over the whole data set, and in situations where algorithm is needed to dynamically adapt to new rules in the data, or when the data are automatically created as a function of time, e.g. stock market price prediction.

The batch learning is a reverse method of online learning. It creates the exact prediction by learning the whole trained data set at once. For example, it analyzes the historical dataset to predict how people lived in previous centuries and the population data of our country to predict the forth coming population survey [8].

III IDENTIFY SUITABLE MACHINE LEARNING METHOD FOR TAMIL EPIGRAPHY DATA ANALYSIS.

In order to compare the machine learning methods mentioned in previous section, this section identifies the suitable machine learning logic for Tamil epigraphy data analysis. The problem in Tamil epigraphy data analysis is how to map language change and identify the structure of the language. Here I review some machine learning methods with some examples, to analyze which one is the most suitable one for Tamil epigraphy data analysis.

Structured Machine learning methods are using rule based analysis to make decisions. Tamil epigraphy corpus analysis contains object marking and Parts of speech (POS) tag database [2]. POS database is used to form rule based formulas for the language and identify its structure and variety. In unstructured machine learning, input is the important factor [9]. In Tamil epigraphy data don't come under any formulas (rule). When we use the method, it has disadvantages while making patterns using the data.

The next method is called semi-structured. It is also a suitable method for Tamil epigraphy data analysis, because it is a combination of structured and unstructured learning [10]. The method can be of use for data which use rules to identify grammar and variety and data which do not come under any rule. Active and passive learning methods are also using rule-based decision making like a structured learning. We apply this method too. Teacher mode machine learning method is needed for Tamil epigraphy because it maintains the rule properly for processing the data, and in the event of any errors occurring while processing, it automatically shows the correct path to process the data for identifying the nature of the language.

An online learning method is not suitable for the process and analyses the Tamil epigraphy data, because it can only access the current online data, not an entire trained dataset. The online data change frequently and Tamil epigraphy contains a vast dataset. The Batch learning method is another suitable method for Tamil epigraphy, because it predicts using the entire trained dataset. This method predicts using the historical data, and since Tamil epigraphy contains a vast number of datasets, the method can predict the grammatical structure of the different varieties of Tamil.

IV. CONCLUSION

This paper has provided a review of the basic methods of machine learning which are suitable for the Tamil epigraphy and data analysis. Machine learning is a huge area in artificial intelligence. It provides various methods to process the data which are automatically thought and done by machines (computers) themselves. It discusses the basic methods for processing data such as structured learning, unstructured learning, semi-structured learning, online and batch learning, teacher mode learning and, passive and active learning. Each of them has its own advantages and disadvantages. For example, structured learning uses rule-based data processing, but the online learning does not depend on the rules, but the current data received for the process. The first one is suitable for Tamil Epigraphy data analysis, while; second one is not. The option of suitable method depends on the type and structure of data used in Tamil epigraphy.

REFERENCES

[1]. Ayon Dey Machine Learning Algorithms: A Review International Journal of Computer Science and Information Technologies, Vol. 7 (3), 2016, 1174-1179

[2]. Diksha Sharma, Neeraj Kumar A Review on Machine Learning Algorithms, Tasks and Applications, International Journal of Advanced Research in Computer Engineering & Technology (IJARCET) Volume 6, Issue 10, October 2017, ISSN: 2278 - 1323

[3]. Appasamy Murugaiyan, Mapping Language Change in Tamil: Corpus analysis and Computer Database Making, In Conference Papers, Tamil Internet Conference, University of Pennsylvania, Philadelphia, June 17-19 2011, 301-306.

[4]. Appasamy Murugaiyan, hero stone inscriptions in tamil (450-650 ce.): text to meaning: a functional perspective* new dimensions in tamil epigraphy:, crea publishers, chennai, pp. 316-351.

[5] Hemant Kumar, Rishabh Choudhary, Comprehensive Review on Supervised Machine

Learning Algorithms, 2017 International Conference on Machine learning and Data Science

[6]. N Pushpa, R Revathi, C Ramya, S Shahul Hameed, Speech processing of Tamil language with back propagation neural network and semi-supervised training, International Journal of Innovative Research in Computer and Communication Engineering 2014 pages;2718-2723, ISSN(Online): 2320-9801 ISSN (Print): 2320-9798

[7]. Bhumika Bhatt, 2 Prof. Premal J Patel, 3 Prof. Hetal Gaudani, A Review Paper on Machine Learning Based Recommendation System, International Journal of Engineering Development and Research, 2014 IJEDR | Volume 2, Issue 4 | ISSN: 2321-9939

[8]. Angra and S. Ahuja, "Machine learning and its applications: A review," 2017 International Conference on Big Data Analytics and Computational Intelligence (ICBDAC), Chirala, 2017, pp. 57-60.

[9]. B Nishanthi, S Shahul Hammed, Detection of Text with Connected Component Clustering, International Journal of Innovative Research in Computer and Communication Engineering, Pages 2434-2440 ISSN(Online): 2320-9801 ISSN (Print): 2320-9798

[10]. Yogesh Singh, Pradeep Kumar, Omprakash Sangwan, A Review Of Studies On Machine Learning Techniques, International Journal of Computer Science and Security, Volume (1) : Issue (1)