Olex a booming Approach for Text categorization

Gandhirajan. A¹, Rajkumar. R¹ and Senthil.R²

Research Scholar, Dept of Computer science, Marudupandiyar College, Thanjavur, Tamilnadu, India.¹
Assistant Professor, Dept of Computer science, Marudupandiyar College, Thanjavur, Tamilnadu, India.²

Abstract: The growth of information is enormous and handling of such information is also become an issue for the researchers. Several algorithms are proposed to handle the information in a better way. OLEX booming law knowledge for Text arrangement, and olex learning problem is stated as an optimization problem relying on the f-measure as the objective function. In this study we focused on biological information/documents, as how this approach could be useful to categorize the molecular information by defining text set and check its arrangements.

Keywords: Text Arrangement; Cluster; Pattern; Threshold Value

I. INTRODUCTION

A text classifier is a program capable of assigning natural language texts to one or more thematic categories on the basis of their contents. A number of machine learning methods to automatically construct classifiers using labeled training data have been proposed in the last few years, including k-nearest neighbors (k-NN), probabilistic Bayesian, neural networks, and SVMs [1-12]. In a different view, rule learning algorithms, have become a successful strategy for classifier induction. Rule-based classifiers provide the desirable property of being interpretable and, thus, easily modifiable based on the user’s a prior knowledge. OLEX, a novel method for the automatic induction of rule-based text classifiers. Here, a classifier is a set of propositional rules, each characterized by one positive literal and (zero or) more negative literals. A positive (respectively, negative) literal is of the form t є d (respectively, ¬ (t є d)), where t is a conjunction of terms t₁^...^ tₙ (a term tᵢ being a n-gram) and d a document [13-17]. Rule induction is based on a greedy optimization heuristics whereby a set of high-quality rules is generated for the category being learned. Unlike other (either direct or indirect) rule induction algorithms, e.g., ripper and c4.5, OLEX is a one-step process, i.e., it directly mines the final rule set, without the need of any post induction optimization [18-24]. In fact, OLEX achieves high performance on all data sets and induces classifiers that are compact and comprehensible. The comparative analysis performed against some state-of-the-art learning methods, notably, NaiveBayes (NB), ripper, c4.5, polynomial SVM, and linear logistic regression (LLR) demonstrates that olex is more than competitive in terms of both predictive accuracy and efficiency [25-33].
II. DESIGN SPECIFICATION

![Flow diagram of Text Categorization](image)

Finding unique words: segmented each corpus into five equalized partitions for cross validation. During each run, four partitions will be used for training, and one for validation. Each of the five combinations of one training set and one validation set is a fold. This process will be useful for finding unique words in input parameters.

Finding occurrence of words: As next process, segmented the corpus into two partitions: the known data and the unknown data. The former is used to induce the model, while the latter is used for testing. As known data are then randomly split into a training set, on which to run algorithm greedy-olex, and a validation set, on which tuning the model parameters. Performed both the above splits in such a way that each category was proportionally represented in both sets. Finally, for every corpus and training set, as scored all terms occurring in the documents of the training set.

Remove least used words using threshold: This module deals with remove the least words using threshold value. Previously, it has calculated the total number occurrence per word. With this calculation help to fix the threshold value for
this process. If does remove the least used word from the database then it has to obtain the most number of occurrence words.

Perform rule based text categorization: It needs to implement the proposed technique called rule based text categorization. This technique has the ability to cluster the number of possible files with the use of most number of occurrence words.

**IV. CONCLUSION**

OLEX presented a novel approach to the automatic induction of rule-based text classifiers. Thus provided a formal description of the method. In particular, described the problem of determining a best set of discriminating terms for a category and proved its intractability. Then, showed how rules are derived from a given set of discriminating terms. Thus, olex predictions require testing the simultaneous presence of several terms along with the simultaneous absence of several sets of terms. While there is in the literature a wide range of rule learning algorithms, one contribution of this project is the form of the hypothesis language. All this makes olex an interesting approach for learning rule-based text classifiers from training sets.

**REFERENCES**

International Journal of Innovative Research in Science, Engineering and Technology  
(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 9, September 2013


Copyright to IJIRSET www.ijirset.com 4382