ISSN: 2319-8753

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 9, September 2013

Olex a booming Approach for Text categorization

Gandhirajan. A[1], Rajkumar. R[1] and Senthil.R[2]

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 ϵ d (respectively, $\neg$ (t ϵ d)), where t is a conjunction of terms t1^.... ^ t_n (a term t_i 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].

ISSN: 2319-8753

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 9, September 2013

II. DESIGN SPECIFICATION

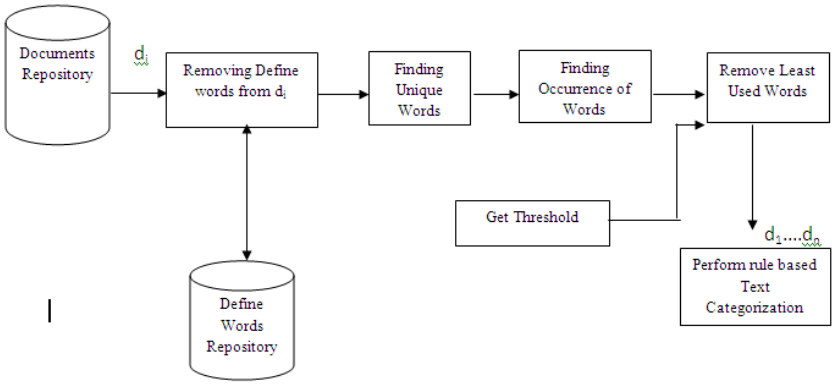

Fig.1 Flow diagram of Text Categorization

III.PROPOSED METHOD

Removing Define words: In this module, algorithm greedy-olex is the search of a "best" classifier over the training set, for given values of the input parameters; the learning process is the search of a "best" classifier over the validation set, for all input parameters values. Then define words removed from documents all words occurring in a list of common stop words, as ill as punctuation marks and numbers. Then, it generated the stem of each of the remaining words, so that documents are represented as sets of word stems.

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

ISSN: 2319-8753

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 9, September 2013

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.

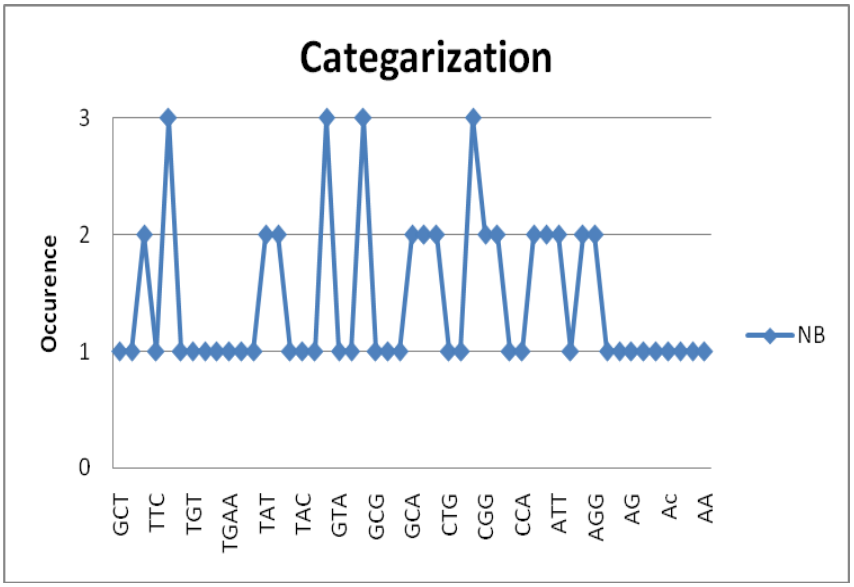

Fig.2 Text Categorization

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

[1] Agresti. A, categorical data analysis. Wiley interscience, 2002.
[2] Anthony.M and Biggs. N, computational learning theory. Cambridge univ. Press, 1992.

ISSN: 2319-8753

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 9, September 2013

[3] Antonie. M and Zaiane. U, "An associative classifier based on positive and negative rules," proc. Ninth acm sigmod workshop research issues in data mining and knowledge discovery (dmkd), 2004.
[4] Apte. C, Damerau. F.J, and Weiss. S.M, "automated learning of decision rules for text categorization," acm trans. Information systems, vol. 12, no. 3, pp. 233-251, 1994.
[5] Baralis. E and Garza. P, "Associative text categorization exploiting negated words," proc. 21st ann. Acm symp. Applied computing (sac '06), pp. 530-535, 2006.
[6] Cohen. W.W, "Text categorization and relational learning," proc. 12th int'l conf. Machine learning (icml), 1995.
[7] Cohen.W and Page.C.D, "polynomial learnability and inductive logic programming: methods and results," new generation computing, vol. 13, no. 34, pp. 369-409, 1995.
[8] Cohen. W.W and Singer. Y, "context-sensitive learning methods for text categorization," acm trans. Information systems, vol. 17, no. 2, pp. 141-173, 1999.
[9] Debole.V and Sebastiani F, "an analysis of the relative difficulty of reuters-21578 subsets," proc. Fourth int'l conf. Language resources and evaluation (lrec '04), 2004.
[10] Dzeroski S, Muggleton S, and Russell S.J, "PAC-learnability of determinate logic programs," proc. Fifth ann. Acm workshop computational learning theory (colt), 1992.
[11] Forman .G, "an extensive empirical study of feature selection metrics for text classification," Machine. J learning research, vol. 3, pp. 1289-1305, 2003.
[12] Gottlob .G, Leone .N, and Scarcello. F, "on the complexity of some inductive logic programming problems," proc.Seventh int'l workshop inductive logic programming (LLP '97), pp. 17-32, 1997.
[13] Hersh.W, Buckley.C, Leone.T, and Hickman. D, "OSHUMED" an interactive retrieval evaluation and new large text collection for research," proc. 17th acm int'l conf. Research and development in information retrieval (sigir '94), Croft. W.B and Van Rijs Bergen. C.J, eds., pp. 192-201, 1994.
[14] Japkowicz. N and Stephen. S, "the class imbalance problem: a systematic study," intelligent data analysis j., vol. 6, no. 5, pp. 429-449, 2002.
[15] Joachims. T, "text categorization with support vector machines: learning with many relevant features," proc. 10th European conf. Machine learning (ecml '98), Ne´dellec. C and Rouveirol. C, eds., pp. 137-142, 1998.
[16] Kietz.J.U and D_zeroski. S, "Inductive logic programming and learnability,"sigart bull., vol. 5, no. 1, pp. 22-32, 1994.
[17] Kloesgen. W, "explora: a multipattern and multistrategy discovery assistant," advances in knowledge discovery and data mining, pp. 249-271, 1996.
[18] Li. W, Han. J, and Pei. J, "cmar: accurate and efficient classification based on multiple-class association rule," proc. First IEEE int'l conf. Data mining (ICDM), 2001.
[19] Pietramala. A, Policicchio. V.L, Rullo. P, and Sidhu I, "a genetic algorithm for text classification rule induction," proc. European conf. Machine learning and principles and practice of knowledge discovery in databases (ecml/pkdd '08), Daelemans. W, Goethals. B, and Morik. K, eds., no. 2, pp. 188-203, 2008.
[20] Quinlan. J.R, "generating production rules from decision trees," proc. 10th int'l joint conf. Artificial intelligence (ijcai '87), pp. 304-307, 1987.
[21] Caropreso M.F, Matwin.S, and Sebastiani. F, "A learner-independent evaluation of the usefulness of statistical phrases for automated text categorization," text databases and document management: theory and practice, Chin. A.G, ed., pp. 78-102, idea group publishing, 2001
[22] Johnson. D.E, Oles. F.J, Zhang. T, and Goetz. T, "A decision-tree-based symbolic rule induction system for text categorization," IBM systems j., vol. 41, no. 3, pp. 428-437, 2002.
[23] Kietz. J.U, "Some lower bounds for the computational complexity of inductive logic programming," proc. Sixth European conf. Machine learning (Ecml '93), vol. 667, pp. 115-123, 1993.
[24] Lewis. D.D, "reuters-21578 text categorization test collection," distribution 1.0, http:/metaxa.net/, 1997.
[25] Open directory project—ODP, http:/dmoz.org, 2008
[26] Sebastiani. F,"Machine learning in automated text categorization," ACM computing surveys, vol. 34, no. 1, pp. 1-47, 2002.
[27] Valiant. L.G, "a theory of the learnable*,"* proc. 16th ann. Acm symp. Theory of computing (stock '84), pp. 436-445, 1984.
[28] Weiss. S and Indurkhya. N, "Optimized rule induction," *ieee expert*, vol. 8, no. 6, pp. 61-69, 1993.
[29 Witten. I.h and Frank. E, data mining: practical machine learning tools and techniques, *seco*nd ed. Morgan kaufmann, 2005.
[30] Wu. X, Zhang.C, and Zhang. S, "Mining both positive and negative association rules," proc. 19th int'l conf. Machine learning '02, pp. 658-665, 2002.
[31] Yang. Y and Pedersen. J.O, "A comparative study on feature selection in text categorization ," proc. 14th int'l conf. Machine learning (ICML '97*)*, Fisher. D.H, ed., pp. 412-420, 1997.
[32] Yang. Y and Liu. X, "A re-examination of text categorization methods*, "* proc. 22nd ACM int'l conf. Research and development in information retrieval ('99), pp. 122-130, 1999.
[33] Cohen, W.W., singer, y.: context-sensitive learning methods for text categorization. ACM transactions on information systems 17(2), 141–173 (1999)