Volume 5, Issue 2, February 2015 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

Novel Technique for Text Data Classification Using Machine Learning

Shashiprabha Singh
PG Scholar, Computer Science & Engineering,
Oriental University, Indore, India

Ashwani Jha
Asst. Professor, Computer Science & Engineering,
Oriental University, Indore, India

Abstract ***This paper presents a novel algorithm based on genetic algorithm and support vectors machine. It use superior quality optimization concert of SVM to get better classification performance of genetic algorithm for Text data classification select dataset and a text dataset are preferred to strength performance of the combing algorithm. It's observably that the hybrid algorithm can be useful to classification in the field of text data categorization. In this study direction will focus on the effect to performance when associated parameter, and proposed framework. Our work includes a comparative study on classification of the same data set with other machine learning algorithms such as support vector machines, genetic algorithm parameters. Since character genetic algorithm SVM parameters to be effective in text categorization, we plan to investigate their efficiency in information retrieval tasks for agglutinative languages.***

Keywords: SVM (Support vector machine), GA (Genetic Algorithm), Text Data Classification, Feature Selection and Text Classification Algorithms.

I. INTRODUCTION

At current, new grown-up and overseas text classification algorithms are decision tree classification algorithm, neural network algorithm, and SVM, KNN classification algorithm. Neural network consists of a collection of neurons; the input unit typically signifies lexical items, the output unit that group or types of attention compute, the weights between neurons communicate terms of dependency. Decision tree is worn as an illustration of the property of the root node, with belongings standards as a branch of the tree. Support vector machines (SVM) when applied to text classification provide excellent precision, but poor recall. One means of customizing SVMs to improve recall, is to adjust the threshold associated with an SVM to suggest an automatic procedure for adjust the thresholds of SVM based GL with enhanced consequences. The level of impenetrability of text classification responsibilities naturally vary. As the number of separate classes increases, so does the impenetrability, and therefore the size of the training set desirable. In any multi-class text classification task, certainly some classes will be added complicated than others to classify. Reasons for this may be: extremely few positive training example for the class, and/or require of first-class predictive features for that class. When training a binary classifier per grouping in text categorization, we use every one the documents in the training quantity that fit in to that category as applicable training data and every the documents in the training corpus that belong to every the other category as non-relevant training data. It is frequently the case that there is an irresistible number of non-applicable training documents particularly when there is a large collection of category with every assigned to a diminutive number of documents, which is characteristically an imbalanced data problem.

This problem presents an exacting challenge to classification algorithms, which can accomplish high accuracy by merely classifying every example as negative. To overcome this problem, cost sensitive learning is desirable. It is the use of in sequence theory attributes the huge number of illustration analyzed and review produce. Frequent technique contains C4.5, CHAID, CART decision tree algorithm, ID3, etc. K-Nearest Neighbor categorization algorithm on behalf of k-nearest neighbor categorization, during the K nearly all comparable history with a permutation to recognize an Innovative Record of Artificial Intelligence.

Neural network technology is mature technology. Neural network consists of a group of neurons, the input unit usually represents lexical items, the output unit that category or categories of interest measure, the weights. SVM were primary recommended by Vapnik and have freshly been used in a range of problems together with model recognition, text categorization and bioinformatics. SVM use kernel occupation to convert input features from minor to higher dimensions [1].

Support vector machine constructs a hyperplane or set of hyper plane in a dimensional space, which can be used for classification, regression or other tasks. Support vector machine categorize data with dissimilar class labels by formative a set of support vectors that are member of the set of preparation inputs that define a hyper plane in the quality space. SVM at hand is a generic method that fits the hyper plane outside to the training data by resources of a kernel function. SVM endeavor to determine a most favorable hyper plane restricted by the input space consequently as to correctly

© 2015, IJARCSSE All Rights Reserved

classify the binary categorization intricacy. The hyper plane is chosen in such a method that there is maximum distance among the hyper plane and the binary examples. The SVM resolve the twofold quadratic programming difficulty to recognize the non-zero Lagrange multipliers and create the most favorable hyper plane. Genetic algorithm is a search algorithm support on the consideration of established genetics. Genetic algorithms have the probable to create collectively the optimal feature subset and SVM parameters at the corresponding time. Our research purpose is to optimize the parameters and characteristic subset concurrently. In this research, we nearby a narrative heuristic text classification technique base on genetic algorithm and support vector machine (SVM).

CHAID is an algorithm. This algorithm selects a set of predictors and their interactions and predicts the optimal value of the dependent variable. In the end what we get is a classification tree. The dependent variable could be a qualitative variable or a quantitative variable.

CART is the ultimate classification tree that has revolutionized the entire field of advanced analytics and inaugurated the current era of data mining. CART, which is continually being improved, is one of the most important tools in modern data mining. Others have tried to copy CART but no one has succeeded as evidenced by unmatched accuracy, performance, feature set, built-in automation and ease of use. Designed for both non-technical and technical users, CART can quickly reveal important data relationships that could remain hidden using other analytical tools.

CART is also used landmark mathematical theory introduced in 1984. Sal ford Systems' implementation of CART is the only decision tree software embodying the original proprietary code. The CART creators continue to collaborate with Salford Systems to continually enhance CART with proprietary advances.

What is decision tree: A decision tree is a tree in which each branch node represents a choice between a number of alternatives, and each leaf node represents a decision. Decision tree are commonly used for gaining information for the purpose of decision -making. Decision tree starts with a root node on which it is for users to take actions. From this node, users split each node recursively according to decision tree learning algorithm. The final result is a decision tree in which each branch represents a possible scenario of decision and its outcome

K nearest neighbors is a simple algorithm that stores all available cases and classifies new cases based on a similarity measure (e.g., distance functions). KNN has been used in statistical estimation and pattern recognition.

Iterative Dichotomiser 3 or ID3 is an algorithm which is used to generate decision tree, details about the ID3 algorithm is in here. There are many usage of ID3 algorithm especially in the machine learning field.

II. BACKGROUND

The use of Support Vector Machines (SVMs) for Text Data Classification for examples. It analyzes the particular properties of learning with text data and identifies why SVMs are appropriate for this task. Because the support vector machine (SVM) has been provide a good performance.

Genetic algorithms are used many increasing application in a variety of search optimization. Machine learning and other problems areas a wide spectrum of disciplines. A general description of GA) are well known techniques widely used for optimization. GA is used Combined with a local search technique are compared to and found to be superior over other methods. This particular GA will be used as the basis for the rest of the paper and it will be enhanced to be able to deal with uncertain release dates and to give as output a solution that is quality robust.

III. RELATED WORK

Wei Xu in at al [1] the major objectives of this learning are double was to create the SVM based collection technique, and to corroborate proposed technique with dissimilar estimate criterion; and how to utilize the proposed all together classifier for credit scoring, and to evaluate classification routine between entity classifiers, and proposed technique. In observation of the two objectives, in this mostly describe the main procedure of building of SVM based company replica, and the purpose of the proposed calculation technique in credit risk assessment, while compare classification performance with dissimilar estimate criteria.

Meijuan Gao in at al [2] they proposed the SVM is a turned in quadratic optimization difficulty, it can guarantee the extremism outcome is the global optimum effect, and it can productively determine the over appropriate problem of ANN. It has high-quality simplification capability and improved classification accurateness. Consequently, the web classification mining technique based on the categorize SVM can augment converge speed and the categorization accuracy to a huge extent. The current classification consequence illustrate that the technique obtainable in this research can accurately categorize the web page content. With the quick progress of WWW, it is a investigate hotspot of in sequence technology by with the web mining technology and organic combine the searching engine to intellectual classify the web page and understand the user individuation service.

Lei Shi in at al [3] proposed a narrative technique for cross lingual text categorization. They have given technique ports a categorization replica qualified in a source language to aim language, with the conversion understanding being learned with the EM algorithm. The replica is extra tuned to fit the allocation in the objective language via semi-supervised learning. Research on dissimilar datasets cover dissimilar languages and dissimilar domains illustrate important improvement over preceding technique that rely on machine translation.

Dilara Torunoğlu in at al [4] in this paper recommend Wikipedia Semantic even technique. Wikipedia is an extremely rich in sequence resource and contain semantic relations such as synonymy, polysemy, hyponymy, associative and definite information and hyperlinks among articles. With WEX we expand our Twitter emotion dataset Wikipedia editorial titles, category and redirects.

© 2015, IJARCSSE All Rights Reserved

IV. PROPOSED METHODOLOGY

In regulate to assess such texts repeatedly. We have developed a number of text classifications systems, such as with fuzzy classification scheme [1], neural networks [2], and support vector machines (SVM). We as well have exposed that appearance collected works technique which is based on direct in organize and a genetic algorithm (GA) are sensible for improving categorization concert. Term com-binations are chosen by GA through two objectives to exploit the number of correctly covert texts on the complex sets and diminish the number of chosen terms. In this research, while a SVM-based text categorization scheme with GA-based term collection is use in reality for assess composed texts, we examine the comparative among the classifications presentation and chosen terms, the implication of texts and preprocessing for classifications with the exploration. Preprocessing consist of remove HTML tags, section word and build Vector Space Model characteristic decrease by rough sets. Our purpose is to discover a lessening with negligible number of attributes, explain. Exchange genotype to phenotype. This phase will exchange every parameter and feature chromosome beginning its genotype into a phenotype. Feature subset later than the genetic process and exchange every feature subset chromosome from the genotype into the phenotype, a feature subset can be resolute. Fitness evaluation for every chromosome instead of C, γ and chosen features, preparation dataset is use to prepare the SVM classifier, as the testing dataset is used to compute classification accuracy. When the classification accuracy is get, every chromosome is assessing by fitness function, termination measure. When the extinction criterion is content, the process ends; or else, we continue with the after that generation of Genetic operation. In this rung, the system searches for enhanced answer by genetic operations, counting selection, mutation, crossover, and substitute. Contribution the preprocessed data sets into the get optimized SVM classifier.

Phase I: following to the text give details the training vector, and articulated as features novel text classification, create Size those at arbitrary to produce the original population.

Phase II: following to type of resemblance and distance, to create the fitness charge of every entity, and assess the fitness of each individual

Phase III: replica to create the population.

Phase IV: Do crossover and mutation procedure in, and choose the training papers in the narrative text with the the popular comparable text.

Phase V: compute being might subsequent phase IV. If the fitness of unit following to Phase IV is improved than the mature one, then exchange the old individual with it;

Phase VI: choose antibodies with huge fitness in the inhabitants to duplicate, and clonally probabilities of antibody character communicate to its fitness.

Phase VII: ensure if the end circumstance is content or not. If it is content, end the algorithm, or else turn to Step II to start the after that iteration.

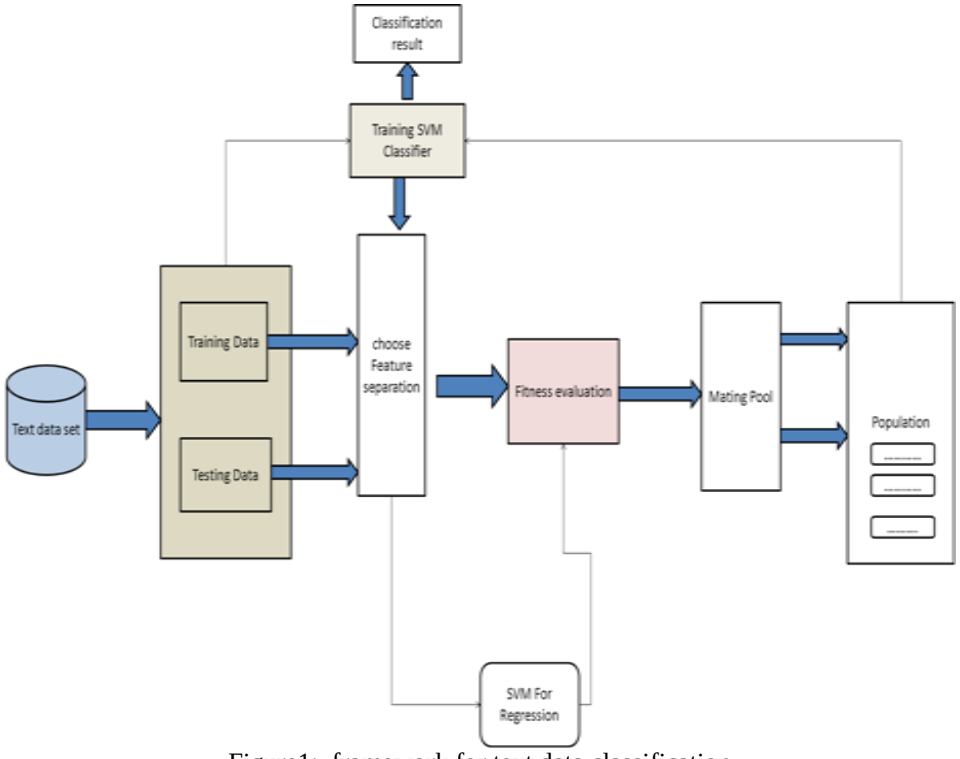

Figure1: framework for text data classification

In this research work, we examine the relative among the classification presentation and chosen terms when SVM-based text categorization system with GA-based term collection is worn really for estimate composed texts. GA-based appearance selection optimizes two objectives which are the max-imagination of properly classified texts and the minimization of chosen terms, and is intend at the expansion in classification presentation on testing sets. Though, the significance of stipulations in the future use was not measured in the entity function. Consequently, still when a number

© 2015, IJARCSSE All Rights Reserved

of conditions have significant role for the categorization, present GA-based expression selection cannot obtain such terms into deliberation. Beginning this motive, the performance deterioration is source in excess of hidden texts in definite use by GA-based term collection since terms which have significant roles for the classification are deleting extremely. In the next description of our classification system, a technique in which terms that will be functional for the classification in the prospect classification are not deleted will be integrated in GA.

In the definite use, unobserved texts that are confidential for improving superiority of have a number of conditions which emerge Initial in those texts. We as well require management such unidentified terms in our after that version.

V. FUTURE WORK

Some problems and concepts that remain unaddressed can be performed in future. Future study direction will focus on the consequence to concert when associated parameters, such as crossing over rate, mutation rate, size of population, etc., have dissimilar values, and get better computational efficiency of the novel algorithm further.

VI. CONCLUSIONS

SVM technique for diminutive sample in the automatic classification has enhanced classification consequences. When the necessitate for a sub-sample to be confidential, when intended to be simply sub-samples and the correspondence of every type of vector that is the internal product, and then choose the class of the record comparison to be sub-sample of the equivalent class. SVM address the diminutive example, nonlinear and high dimensional pattern recognition presentation of a lot of unique compensation, and can be functional to purpose estimate and extra machine learning problems.

ACKNOWLEDGMENT

We would like to express our gratitude to all those who gave us the possibility to complete this paper. We want to thank the computer science department of the Oriental University, Indore for giving me permission to commence this paper in the first instance, to do the necessary research work and to use departmental data. We are deeply indebted to our supervisor from the CS Department whose help, stimulating suggestions and encouragement

REFERENCES

[1] Wei Xu, Shenghu Zhou, Dongmei Duan, Yanhui Chen," A Support Vector Machine Based Method For Credit Risk Assessment" IEEE International Conference on E-Business Engineering978-0-7695-4227-0/10 2010 IEEE.

[2] Meijuan Gao, Jingwen Tian, Shiru Zhou," Research of Web Classification Mining Based on Classify Support Vector Machine" ISECS International Colloquium on Computing, Communication, Control, and Management-2009.

[3] Lei Shi, Rada Mihalcea, Mingjun Tian," Cross Language Text Classification by Model Translation and Semi-Supervised Learning" Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, pages 1057–1067.

[4] Dilara Torunoğlu, Gürkan Telseren, Özgün Sağtürk, Murat C. Ganiz," Wikipedia Based Semantic Smoothing for Twitter Sentiment Classification" 978-1-4799-0661-1/13/-2013 IEEE.

[5] Shitao Zhang Xiaoming Jin Dou Shen Bin Cao Xuetao Ding Xiaochen Zhang," Short Text Classification by Detecting Information Path" CIKM'13, Oct. 27–Nov. 1, 2013, San Francisco, CA, USA.

[6] X. Wan.. Co-training for cross-lingual sentiment classification. In Proceedings of the Joint Conference of the Association of Computational Linguistics and the International Joint Conference on Natural Language Processing, Singapore, August.- 20091067

[7] C. Banea, R. Mihalcea, J. Wiebe, and S. Hassan. Multilingual subjectivity analysis using machine translation. In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP 2008), Honolulu, Hawaii-2008.

[8] A. Gliozzo and C. Strapparava. Exploiting comparable corpora and bilingual dictionaries for crosslanguage text categorization. In Proceedings of the Conference of the Association for Computational Linguistics, Sydney, Australia. 2006.

[9] Li, S.T., Wu, X.X., and Hu, X.Y.: 'Gene selection using genetic algorithm and support vectors machines', Soft Computing, 2008, 12, (7), pp. 693-698.

[10] Kim, D.S., and Park, G.S.: 'Modeling network intrusion detection system using feature selection and parameters optimization', Ieice Transactions on Information and Systems, 2008, E91D, (4), pp. 1050-1057.

[11] Go, A., Bhayani, R., & Huang, L. (2009). Twitter sentiment classification using distant supervision. CS224N Project Report, Stanford, 1-12.

[12] Wang, P., Hu, J., Zeng, H. J., & Chen, Z. (2009). Using Wikipedia knowledge to improve text classification. Knowledge and Information Systems,

[13] Joachims, T.: Text categorization with support vector machines: learning with many relevant features. In: N´edellec, C., Rouveirol, C. (eds.) ECML 1998. LNCS,vol. 1398, pp. 137–142. Springer, Heidelberg (1998).

[14] ZHONG Jiang, WEN Luo-Sheng, FENG Yong, YE Chun-Xiao and LI Zhi-Gu. Study on the Web lassification Based on Proximal Support Vector Machine. Computer Science, 2008, 35 (3), pp.167-169,202.

© 2015, IJARCSSE All Rights Reserved

[15] Liu, S., Jia, C.Y., and Ma, H.: 'A new weighted support vector machine with GA-based parameter selection', Proceedings of 2005 International Conference on Machine Learning and Cybernetics, Vols 1-9, 2005, pp. 4351-4355

[16] ZHANG X R, LIU F. A patten classification method based on GA and SVM. 6th International Conference on Signal Processing Proceedings, Vols I and Ii, 2002, pp.110-113.

[17] F. Sebastiani, "Text categorization", Alessandro Zanasi (ed.) Text Mining and its Applications, WIT Press, Southampton, UK, pp. 109-129, 2005

[18] A. Dasgupta, P. Drineas, B. Harb, "Feature Selection Methods for Text Classification", KDD'07,ACM, 2007.

[19] N. E. Ayat, M. Cheriet, and C. Y. Suen, "Automatic model selection for the optimization of SVM kernels," S. Keerthi and C.-J. In Asymptotic Behaviors of Support Vector Machines with Gaussian Kernel," Neural Computation, vol. 15, pp. 1667-1689, 2003.

© 2015, IJARCSSE All Rights Reserved