ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 8, August 2015

A Study on Web Page Classification using Machine Learning Algorithms

B.Sundarraj, K.P.Kaliyamurthie, Sundararajan.M, Arulselvi S

Assistant Professor, Department of Computer Science Engineering, Bharath University, Chennai, India

Professor & Head, Department of CSE, Bharath University, Chennai, India

Director, Research Center for Computing and Communication, Bharath University, Chennai, Tamil Nadu, India

Co-Director, Research Center for Computing and Communication, Bharath University, Tamil Nadu, India

ABSTRACT : In this paper we tend to use machine learning algorithms like SVM, KNN and GIS to perform a behavior comparison on the net pages classifications drawback, from the experiment we tend to see within the SVM with tiny range of negative documents to make the centroids has the littlest storage demand and also the least on line take a look at computation value. however most GIS with completely different range of nearest neighbors have a fair higher storage demand and on line take a look at computation value than KNN. this means that some future work ought to be done to do to cut back the storage demand and on list take a look at value of GIS.

KEYWORDS: net Classifications, Machine Learning, LIBSVM, SVM, K-NN.

I.INTRODUCTION

Nowadays, the net pages is growing at Associate in Nursing exponential rate and may cowl virtually any info required. However, the massive quantity of websites makes it additional and tougher to effectively realize the target info for a user. usually 2 solutions exist, class-conscious browsing and keyword looking. However, these pages vary to a good extent in each the knowledge content and quality. Moreover, the organization of those pages doesn't yield straightforward search. therefore Associate in Nursing economical and correct methodology for classifying this Brobdingnagian quantity of knowledge is incredibly essential if the net pages is to be exploited to its full potential. This has been felt for a protracted time and plenty of approaches are tried to resolve this drawback. many alternative Machine learning- primarily based algorithms are applied to the text classification task, together with k-Nearest Neighbour (k-NN Algorithm) [2], Bayesian rule [7], Support Vector Machine (SVM) [5], Neural Networks [8], and call trees [9].

In past years, there have been in depth investigations and speedy progresses in mechanically class-conscious classification. Basically, the models rely on the hierarchical data structure of the dataset and also the assignment of documents to nodes within the structure. withal, most of researches adopt a top-down model for classification. M.-Y. Kan and H. O. N. Thi. [9]; have summarized four structures employed in text classification: virtual class tree, class tree, virtual directed acyclic class graph and directed acyclic class graph. The second and forth structures ar popularly employed in several net directory services since the documents may well be assigned to each internal and leaf classes. These 2 models used category-similarity measures and distance-based measures to describe the degree of incorrectly classification in decision making the classification performance. F. Sebastiani [5]; found a lift of the accuracy for class-conscious models over flat models employing a consecutive Boolean call rule and a increasing call rules. They claimed solely few of the comparisons ar needed in their approaches attributable to the potency of consecutive approach. They even have shown that SVM take an honest performance for virtual class tree, however class tree isn't thought of in their work. E. Gabrilovich and S. Markovitch [6]; planned Associate in Nursing economical optimisation rule, that relies on progressive conditional gradient ascent in

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 8, August 2015

single-example sub-spaces spanned by the marginal twin variables. The classification model may be a variant of the utmost Margin Markoff Network framework, that is provided with Associate in Nursing exponential family outlined on the perimeters. This methodology resolved the scaling drawback to medium-sized informationsets however whether or not it's suited Brobdingnagian quantity data set is unclear. N. aeronaut and N. Nguyen[2]; used a refined analysis, that turns the class-conscious SVM classifier into Associate in Nursing approximate price of the Thomas Bayes best classifier with relevancy a straightforward random model for the labels. They proclaimed Associate in Nursing improvement on class-conscious SVM rule by exchange its top-down analysis with a algorithmic bottom-up theme. Loss operate is employed to live the performance in classification filed during this paper. The aim of our work is to debate the matter of unbalanced information and measuring of multi-label classification and realize ways to resolve them. In the experiments, we tend to compare the classification performance victimisation each the quality measurings precision/recall and extended measurement loss price.

II.MACHINE LEARNING ALGORITHMS

2.1. Support Vector Machines (SVM)

Support Vector Machine is Associate in Nursing machine learning technique supported applied math Learning theory. SVM has been evidenced to be terribly effective in handling high-dimensional feature areas, the foremost difficult drawback of alternative machine learning techniques as a result of the alleged curse of spatiality. For simplicity, let's examine the essential plan of SVM within the linear- severable case (i.e., the coaching sample is severable by a hyper plane). supported Structural Risk diminution principle, SVM rule try and realize a hyper plane specified the margin (i.e., the smallest distance of any coaching purpose to the hyper plane,) is perfect. so as to seek out Associate in Nursing best margin, quadratic optimisation is employed within which the essential computation is that the real of 2 points within the input house. For nonlinear-separable case, SVM use kernel-based ways to map the input house to a alleged feature house. the essential plan of kernel ways is finding a map Φ from the input house that is nonlinear severable to a linear-separable feature house ; two. However, the matter with the feature house is that it's typically of terribly massive or perhaps infinite dimensions and therefore computing real during this house is stubborn. fortuitously, kernel ways overcomes this drawback by finding maps specified computing dot product in feature areas becomes computing kernel functions in input areas
k(x,y) = <Φ(x).Φ(y)>;
where k(x,y) may be a kernel operate within the input house and <Φ(x).Φ(y)> may be a real within the feature house. Therefore, the real in feature areas is computed although the map Φ is unknown. Some most generally used kernel functions ar Gaussian RBF, Polynomial, Sigmoidal, and B-Splines.

2.2. K-Nearest Neighbour

In these ways, there's a family of alleged Memory primarily based Learning or instance-based learning algorithms like kNN and Rocchio. In these algorithms, in distinction to learning ways that construct a general, express description of the target operate once coaching examples ar provided. Those algorithms merely store the coaching examples as they're or remodel the coaching examples into some alternative objects before store them. All those objects ar referred to as instances or cases. once when a brand new question instance is encountered, its relationship to any or all the previous keep instances is calculated so as to assign a target operate price for the new instance. this sort of rule usually includes a low off line computation coaching value however a high storage demand and a high on line take a look at computation value. In most instanced primarily based learning ways, they use a typical technique to represent a document. every document may be a feature vector of the shape ,here di is that the ith feature of the document and m is that the range of options. Typically, m are the vocabulary size and every feature represents a selected word. completely different coefficient schemes to work out the weights are introduced and most of them embrace binary weight, term frequency and combined with inverse document frequency. generally these weights ar normalized by divided by issue|an element} of the factor norm. K-nearest neighbor (kNN) alogirthm is one among the instance-based learning algorithms. knowledgeable Network [3]; is understood collectively of best text categorization classifier designed with this rule. within the coaching section, every document is preprocessed to a vector as descript before

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 8, August 2015

and keep within the instances assortment. within the on-line coaching section, the similarity between question document Associate in Nursing every instance is calculated as: The score of documents belong to each corresponding class is accumulated and a threshold is set to assign some classes to the present new unknown document. solely the score of the k nearest neighbor documents of this new unknown document question is taken into consideration.

2.3. Generalized Instance Set (GIS) Algorithm:

therefore is there some methodology that may mechanically modify the coarseness of the coaching instance set size Associate in Nursingd keep SVM and kNN's blessings the maximum amount as possible? Generalized Instance Set (GIS) [4]; is such an rule for this concept. GIS by selection substitutes applicable positive and negative coaching documents within the coaching document pool to get rid of some noise within the coaching set. It mechanically selects a representative positive instance and performances a generalization. really it will this generalization procedure in an exceedingly supervised manner. It brings out Associate in Nursing analysis metric to live the illustration ability of a generalized instance set and will the generalization steps in an exceedingly greedy to extend the illustration ability. The representative operate Rep(G) for a generalized instance G is outlined as follows:
The pseudo-code is as follows:

- Input: The coaching set T
- The class C
- Procedure GIS(T,C)
- Let G and G_new be generalized instances.
- GS be generalized instance set, and GS is initialized to empty.
- Repeat
- Randomly choose a positive instance as G.
- Rank instances in T in keeping with the similarity
- Compute Rep(G)
- Repeat
- G_new=G
- G=Generalize(G_new,k nearest neighor)
- Rank instances in T in keeping with the similarity
- Compute Rep(G)
- Until Rep(G)
- Rank instances in T according to the similarity
- Compute Rep(G)
- Until Rep(G)<Rep(G_new)
- Add G_new to GS
- Remove top k instances from T.
- Until no positive instances in T
- Return GS

In this algorithm, we use a SVM formula that is described in previous section for the generalization step.

III. TRAINING PHASE

The SVM classifier of our application is implemented based on the LIBSVM library. The library realizes C-support vector classification (C-SVC), ν-support vector classification (ν-SVC), ν-support vector regression (ν-SVR), and incorporates many efficient features such as caching, sequential minimal optimization and performs well in moderate-sized problems (about tens of thousands of training data points). In our application, we only use C-SVC with RBF kernel function To train our SVM classifier we use a collection of 7566 Aljazeera News web documents that are already categorized in 16 broad categories by

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 8, August 2015

domain experts. After parsing and pre- processing, the total number of terms are 189815, meaning that each document is represented by a 189815-dimensional sparse vector – a large dimension that is considered to be extremely difficult in other machine learning techniques such as neural networks or decision trees. Aside from the training data set, another testing data set consisting of 490 web documents that are also already classified in the same 16 categories is used to evaluate the classifier. To decide the best parameters (i.e., C and γ) for the classifier, a grid search (i.e., search over various pairs (C, γ)) is needed to be performed. However, since a full grid search is very time-consuming we fix the γ parameter as 1/k (k is the number of training data) and only try various values of C. A recommended range of values of C is $2^0, 2^1, 2^2, \ldots, 2^{10}$ which is known good enough in practice.
the number of nearest neighbors as 40.

IV.RESULT ANALYSIS

4.1 Negative Documents in SVM:

There have been some arguments for the effectiveness of using negative documents in constructing the centroids for the GIS algorithm. Some people c
laimed that using a rather small number of negative is enough for GIS. Some experiments are done for this problem. In the following graph, the x-axis represents the number of negative documents used to build the centroids. The y-axis is the performance of GIS algorithm that is tried with 10-fold cross validation. We can see that the performance of SVM algorithm does has relationship with the number of negative documents used to build the centroid in the corpus. And it is interesting to see that precision seems to increase monotonically with the number of negative documents. It can be explained intuitively, as more and more negative documents are subtracted from the centroid, the non-relevant document will have a larger distance with the centroid so the precision will get higher. But the recall's trend is more complex.

4.2 The Number of K nearest neighbors in GIS

In the following graph, the x-axis represents the number of k nearest neighbors used to build the generalized instance, and y-axis represents the performance of GIS algorithms by 10- fold cross validations. We can see from the graph that there is a tradeoff between the number of nearest neighbors and the performance. Intuitively speaking, a too small number of nearest neighbors cannot give a very much generalization power to the centroid and cannot make the centroid insensitive to the noise too much. But too much number of the nearest neighbors will ignore the local detail information. So there should be such kind of trade off in the number of near neighbors.

V.CONCLUSION

From the graphs, we can see SVM with small number of negative documents to build the centroids has the smallest storage requirement and the least on line test computation cost. But almost all GIS with different number of nearest neighbors have a even higher storage requirement and on line test computation cost than KNN. This suggests that some future work should be done to try to reduce the storage requirement and on list test cost of GIS. But in reality, people often only keep most important (high weight) terms in each centroid instead of keeping everything. So in that case, the analysis does not make much sense. But if we cut off the number of terms, the performance will be influenced. There is also some tradeoff for this factor. And that is something that we should do more work to investigate. GIS is some idea that wants to find the true granularity for instances' set in an automatically and supervised way. There was some other work that tries to do the same work in unsupervised way [5]. Their results are different on two different Datasets. It seems that our experiment does not support the effectiveness of this idea. But there needs more work to explore this work.

REFERENCES

[1] E. Baykan, M. Henzinger, L. Marian, and I. Weber. A comprehensive study of features and algorithms for url-based topic Classification. ACM Transactions on the Web, 5:15:1–15:29, July 2011.

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 8, August 2015

[2] P. N. Bennett and N. Nguyen. Refined experts: improving Classification in large taxonomies. In Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval, pages 11–18. ACM, 2009
[3] Jayaraman B., Valiathan G.M., Jayakumar K., Palaniyandi A., Thenumgal S.J., Ramanathan A., "Lack of mutation in p53 and H-ras genes in phenytoin induced gingival overgrowth suggests its non cancerous nature", Asian Pacific journal of cancer prevention : APJCP, ISSN : 13(11) (2012) PP. 5535-5538.
[4] A. Broder, M. Fontoura, E. Gabrilovich, A. Joshi, V. Josifovski, and T. Zhang. Robust Classification of rare queries using web knowledge. In Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pages 231–238, New York, NY, July 2007. ACM Press
[5] C. Castillo and B. D. Davison. Adversarial web search. Foundations and Trends in Information Retrieval, 4(5):377–486, 2010.
[6] Kalaiselvi V.S., Saikumar P., Prabhu K., Prashanth Krishna G., "The anti Mullerian hormone-a novel marker for assessing the ovarian reserve in women with regular menstrual cycles", Journal of Clinical and Diagnostic Research, ISSN : 0973 - 709X, 6(10) (2012) PP.1636-1639.
[7] F. Sebastiani. Machine learning in automated text categorization. ACM Computing Surveys, 34(1):1– March 2002.
[8] E. Gabrilovich and S. Markovitch. Text categorization with many redundant features: Using aggressive feature selection to make SVMs competitive with C4.5. In Proceedings of the 21st International Conference on Machine Learning, page 41, New York, NY, 2004. ACM Press.
[9] Subhashree A.R., Shanthi B., Parameaswari P.J., "The Red Cell Distribution Width as a sensitive biomarker for assessing the pulmonary function in automobile welders- a cross sectional study", Journal of Clinical and Diagnostic Research, ISSN : 0973 - 709X, 7(1) (2013) PP. 89-92.
[10] H. Guan, J. Zhou, and M. Guo. A class-feature-centroid classifier for text categorization. In 18th International World Wide Web Conference, pages 201–201, April 2009.
[11] C.-C. Huang, S.-L. Chuang, and L.-F. Chien. Liveclassifier: Creating hierarchical text classifiers through web corpora. In Proceedings of the 13th International Conference on World Wide Web, pages 184–192, New York, NY, 2004. ACM Press
[12] Gopalakrishnan K., Prem Jeya Kumar M., Sundeep Aanand J., Udayakumar R., "Analysis of static and dynamic load on hydrostatic bearing with variable viscosity and pressure", Indian Journal of Science and Technology, ISSN : 0974-6846, 6(S6) (2013) PP.4783-4788.
[13] M.-Y. Kan and H. O. N. Thi. Fast webpage Classification using URL features. In Proceedings of the 14th ACM International Conference on Information and Knowledge Management, pages 325–326, New York, NY, 2005. ACM Press.
[14] H. Malik. Improving hierarchical svms by hierarchy flattening and lazy Classication. In Proceedings of Large-Scale Hierarchical Classification Workshop, 2010.
[15] Srinivasan V., "Analysis of static and dynamic load on hydrostatic bearing with variable viscosity and pressure", Indian Journal of Science and Technology, ISSN : 0974-6846, 6(S6) (2013) PP.4777-4782.
[16] X. Qi and B. D. Davison. Hierarchy evolution for improved Classification. In Proceedings of the 20th ACM International Conference on Information and Knowledge Management, pages 2193–2196, October 2011.
[17] W. A. Awad, S. M. ELseuofi ; "Machine Learning Methods for Spam E-Mail Classification"; Int. J. of Computer Applications (IJCA); Vol. 16; No. 1; 2011.
[18] Dr.A.Muthu Kumaravel, Data Representation in web portals, International Journal of Innovative Research in Computerand Communication Engineering, ISSN(Online): 2320-9801,pp 5693-5699, Vol. 2, Issue 9, September 2014
[19].Dr.Kathir.Viswalingam, Mr.G.Ayyappan,A Victimization Optical Back Propagation Technique in Content Based Mostly Spam Filtering ,International Journal of Innovative Research in Computerand Communication Engineering ,ISSN(Online): 2320-9801 , pp 7279-7283, Vol. 2, Issue 12, December 2014
[20].KannanSubramanian,FACE ECOGNITION USINGEIGENFACE AND SUPPORT ECTORMACHINE,International Journal of Innovative Research in Computerand Communication Engineering,ISSN(Online): 2320-9801,pp 4974-4980, Vol. 2, Issue 7, July 2014.
[21].Vinothlakshmi.S,To Provide Security & Integrity for StorageServices in Cloud Computing ,International Journal of Innovative Research in Computer and Communication Engineering ,ISSN(Online): 2320-9801 , pp 2381-2385 ,Volume 1, Issue 10, December 2013