International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878, Volume-8 Issue-5, January 2020

Support Vector Machine for Text Categorization using Principle Component Analysis in Data Mining

A. Ravi Kumar, G. Anil Kumar

Abstract: ***Data mining is the withdrawal of concealed prescient information also obscure data, examples, connections and learning by investigating the enormous informational collections which are hard to discover and distinguish with customary measurable techniques. The major issues in text categorization are classification accuracy and computation time. To overcome these issues, an efficient classification method is needed for high differentiation exactness as fine as minimizing the computation period. In this work, we propose the classification of data using support vector machine for text categorization along with principle component analysis. Bolster Vector Machines is a managed learning system with numerous attractive characteristics that make it a prevalent calculation. Principle Component Analysis (PCA) is the feature removal technique is used towards mine the features with in the text. Chi-Square is a further assortment technique it is used to selecting the features from removed features. Finally by this proposed work, the classification accuracy also computation period is improved than other existing algorithms in many applications.***

Keywords: Principle Component Analysis (PCA), Chi-Square.

I. INTRODUCTION

The quick advancement of AI innovations gives us a decent opportunity to process and break down information in a pristine point of view. AI, otherwise called information revelation, is one of the most significant parts of software engineering, which intends to discover helpful examples from information and is very unique in relation to those conventional factual techniques. As a nearly new AI calculation, Support Vector Machine (SVM) has pulled in much consideration as of late and has been effectively utilized in different application jobs [1–6]. In this investigation, we center around developing an enhanced SVM model, in order to utilize it on coronary illness information grouping, intending to improve the characterization effectiveness and precision of SVM.

Manuscript received on January 02, 2020.
Revised Manuscript received on January 15, 2020.
Manuscript published on January 30, 2020.
* Correspondence Author
A. Ravi Kumar*, Research scholar, SriSatya Sai University of Technology & Medical Sciences, Sehore, Bhopal, Madhya Pradesh,
G.Anil Kumar, Principal, Scient Institute of Technology, Hyderabad, Telangana.

II. RELATED WORK

Numerous sorts of writing have included the substance of utilizing Support Vector Machine toward manage information.

Muthu Rama-Krishnan/et-al. structured an SVM founded classifier, which remained utilized in 2 UCI mammogram datasets on behalf of bosom malignant growth location and arrived at the precision of 99.385% and 93.726%, individually [7]. Xie and Wang incorporated a crossbreed include determination technique
with SVM for them at squamous malady conclusion, which arrived at the exactness of 98.61% [8]. Highlight choice is the premise of AI calculations; fitting component determination methodology can clearly improve the exhibitions of AI techniques. Deisy et al. planned a novel data hypothesis founded component choice calculation to develop the characterization exactness on behalf of SVM classifiers/[9–12]. Additional component choice strategies, for example, common data estimation [13], portion *F*-score highlight choice, and express edge-based element end strategy are frequently embraced to show signs of improvement grouping results for SVM or other AI calculations [14–17]. The vast majority of the AI calculations consume their limitations; legitimate procedures ought to be occupied to choose the streamlined estimations of them. Hereditary Algorithm, Particle/Swarm Optimization Algorithm, Artificial/Immune System Algorithm, also Grid Search Techniqueremain those frequently utilized limitation streamlining calculations [18–23]. For the most part, information highlight determination techniques and parameter improvement procedures are extensively measured. Lin et al. built up a ReplicatedAnnealingmethodon behalf oflimitation assurance also highlight choice in SVM, in addition to tests demonstrated the great execution of it [24]. Tan et al. projected another half and half approach, in which Genetic Algorithm and Support.

Vector Machine are incorporated adequately dependent on a wrapper strategy[25], which performed well on the UCI chromosome dataset. Literature[26] displayed a mixture approach dependent on highlight choice, fluffy weighted preprocessing, and fake invulnerable acknowledgment framework, which was utilized on the UCI coronary illness and hepatitis ailment datasets, and the got exactnesses are 92.39% and 81.82%, separately.

Other than highlight choice and parameter advancement, a bit capacity is another factor that ought to be considered for the part based AI calculations like SVM. Khemchandani et al. received an ideal part determination strategy in Twin Support Vector Machines; its proficiency was affirmed with some UCI AI benchmark datasets [12, 27]. Abibullaev et al. presented a direct programming SVM with multi-bit work for mind signal information grouping and got a decent execution [15,28–30]. Counterfeit Neural Network [23], Extreme Learning Machine [31], *K*-Nearest Neighbor investigation, Fuzzy Logic-based strategies [32–34], Ensemble Learning calculations [34–39],

etc, are frequently utilized or crossbreed used to complete information arrangement undertakings and more often than not can get great characterization results. As another variant of Support Vector Machine, Least Square SVM includes fairness limitations rather than imbalance imperatives alsoworkingsthrough a least-squares price work. A conspicuous downside of the Least Square SVM is that the fight faculties are missing [22,40,41]. Yang-et-al .built up a versatile pruning calculation dependent on the base to-top system, in which the gradual alsodecrementalknowledge techniques stood utilized also settled the disadvantage of conventional Least/Square -SVM[42–45]. Concluded the examination of prevailing writing, we saw that the central matters of studies utilizing Support Vector Machine for grouping are changing also using ordinary arrangement calculations in the mix and attempting to get better order execution. All in all, the primary procedure incorporates three techniques: (1) Data Preprocessing (highlight choice, standardization, measurement decrease, etc.);(2) Constructing Optimized Classification Models (counting limitation enhancement); (3) Classification Accuracy also Efficiency Demonstration. Albeit heaps of endeavors consume been completed on SVM in addition to its submissions, its exhibition is unwanted and still should be improved.

III. IMPLEMENTATION

The planned System Regularization, Stratified Cross Authentication,and Multi level Grid Search founded SVM algorithm remain described in the Proposed Methods part; equivalentinvestigatesremainpresented in the InvestigationalOutcomes part; the subsequentportionprovides the finishes of this study.

PCA is a traditional measurable technique for changing characteristics of a dataset hooked on another arrangement of uncorrelated traits named head parts (PCs). PCA container be utilized to diminish the dimensionality of a dataset, though as yet holding however much of the inconstancy of the dataset as could be expected. High dimensional information can act issues for AI like prescient models dependent on such information risk overfitting. Moreover, a considerable lot of the properties might be excess or profoundly connected, which can likewise prompt a corruption of expectation exactness. There are numerous instances of the utilization of AI to group high dimensional information, for example, quality articulation microarray information [1], picture information [2] and content order [3]. Another case of high dimensional information, ghastly information, is utilized for the investigations displayed in this paper. In the order undertaking considered here, Raman spectra might be utilized for the programmed distinguishing proof of a substance inside a material. Regularly, techniques commencing a field of concentrate recognized as chemo metrics consume been applied to this specific issue [4], and these strategies use PCA to deal with the great dimensional spectra. The objective of this examination is to decide whether PCA containerremain utilized to develop the presentation of AI techniques in the characterization of such great dimensional information.

In the accompanying portrayal, the dataset is spoken to through the lattice X, where X is a N × p grid. For phantom requests, every line of X, the p-vector xi comprises the forces at every wavelength of the range test I. Every section, Xj contains every one of the perceptions of one quality. PCA is utilized to beat the recently referenced issues of high-dimensionality also co-linearity through lessening the quantity of indicator properties. PCA changes the arrangement of information sources X1, X2, . . . , XN into additional arrangement of segment vectors T1, T2, . . . , TN where the T 's consume property that a large portion of the first statistics data substance (or the greater part of its change) is put away in the initial couple of T 's (the key segment scores). The thought is this enables decrease of the information to fewer measurements, with uninformed misfortune, just by disposing of a portion of the primary parts (PCs). Every PC is a direct blend of the first inputs and every PC is symmetrical, which along these lines wipes out the issue of co-linearity. This direct change of the framework X is determined by a p × p network P with the goal that the changed factors T are assumed by:

T = XP or alternatively X is decayed as follows: X = T P T

Where P is recognized as the loadings network. The sections loadings network P can be determined as the eigenvectors of the grid XT X [8], a figuring which container be computationally concentrated when managing datasets of 500-3000 qualities.

Text Classifiers:

There are many classifiers that have been developed for variety of tasks in text classification and they give acceptable accuracy. Among them we will use the following classifiers and show how the accuracy improved after using PCA as a feature extraction:

- Random Forest (RF): RF is a very good, powerful, robust and versatile learning technique, however it is a promise choice for high-dimensional text data. It is introduced in 2000s [4], it is a popular classification method which builds multiple decision trees (not only one), which are utilized to decide the ultimate result. For grouping issues, the outfit of basic trees votes in favor of the most well-known class. One of the most known woodland development techniques, proposed by Breiman, is a subspace of highlights which are picked haphazardly at every hub to develop parts of the choice trees, at that point sacking strategy is utilized to produce preparing information subsets for structure individual trees, at last mix of every single individual tree are framed to shape irregular timberlands model [4].

• Support Vector Machine (SVM): SVM has stoodaccepted as single of the greatest effective text categorization method. It gives high classification accuracy especially in highly dimensional data such that it controls complexity and over fitting issued. The time taken for each process is less than the other classifiers. So that it becomes an acceptable choice for large data set as textual data.

SVMs are designed to handle high-dimensional data. SVM was developed in 1995 by Cortes and Vapnik [22]. Its core idea behind SVM is to invention an optimum hyper plane between sets of hyper plane

that maximize hyperplane edge, which is the good ways from the hyperplane to the closest purpose of the example [22] [23].

The report delegates which are nearest to the chosen surface are known as the help vectors. SVM is fundamentally used to amplify the edge, which will guarantee that the info example would be ordered effectively [24]. The point of SVM is to discover the most ideal grouping capacity so as to separate between individuals from two classes in the preparation information in a two-class learning task.

• Choice Tree Algorithm (J-48): The choice tree reproduces the manual arrangement of preparing archives by making very much characterized genuine/false-inquiries as a tree arrangement. In the choice tree structure, leaves speak to the relating classification of archives also branches speak to conjunctions of highlights that lead to those classes. The tree grows until every single content is sorted accurately or inaccurately. The efficient choice tree can without much of a stretch categorize a record through placing it in the root hub of the tree also let it go concluded the inquiry structure awaiting it arrives at a specific leaf which speaks to the objective on behalf of the arrangement of the archive. The choice tree order technique has a few points of interest over other choice help apparatuses. The principle favorable position of a choice tree is that it is simple incomprehension and translating, notwithstanding for non-balanced clients. Likewise, they are a power to loud information and they can learn disjunctive articulations appear to be appropriate for content arrangement. The significant disadvantage of utilizing a choice tree is completed fits the preparation information through the event of an elective tree that sorts the preparation information more awful however would arrange the archives to be ordered better. One of the most notable choice tree calculations is J-48 that we will use in our work.

IV. EXPREIMENTS AND RESULTS

In our study, subsequently pre-processing the data by eliminating the end words which are worthless for classification. In this section, we existent a series of investigates for text classification on three different standards and popular text collection to examine the performance of using principle components analysis (PCA) with different text classifiers; standard RF, standard SVM.

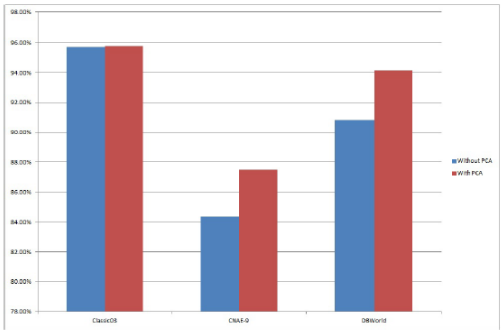

Fig.1. Results with PCA and with out PCA classification.

V. CONCLUSION

In this study, we use principle/component/ analysis/(P/C/A) as a feature removal method to decrease the great dimensionality of the dataset. PCA eliminates the irrelevant, noisy and redundant parameters commencing the feature vector also thus develop the presentation of the classifier for text classification. Major, we pre-process the documents where the feature vector is attained over dissimilar steps similar stop words elimination, stopping also indexing.Then the classifiers we mentioned above are applied on the extracted features. Most of the obtained consequences demonstration that the presentation of most of the classifiers improved after using PCA and this seems very promising for text categorization applications. The results has proved the methods used can filter the data by eliminating the concessionary data along with good hyperactive data.

REFERENCES

1. N. Ur-Rahman and J. A. Harding, “Textual data mining for industrial knowledge management and text classification: A business oriented approach,” Expert Systems with Applications, vol. 39, no. 5, pp. 4729– 4739, 2012.
2. Y. Yang and J. O. Pedersen, “A comparative study on feature selection in text categorization,” in Proceedings of the Fourteenth International Conference on Machine Learning, ICML ’97, (San Francisco, CA, USA), pp. 412–420, Morgan Kaufmann Publishers Inc., 1997.
3. T. Joachims, “Text categorization with support vector machines: Learning with many relevant features,” Machine learning: ECML-98, pp. 137– 142, 1998.
4. L. Breiman, “Random forests,” Machine learning, vol. 45, no. 1, pp. 5–32, 2001.
5. M. Ghiassi, M. Olschimke, B. Moon, and P. Arnaudo, “Automated text classification using a dynamic artificial neural network model,” Expert Systems with Applications, vol. 39, no. 12, pp. 10967–10976, 2012.
6. M. Zareapoor and K. Seeja, “Feature Extraction or Feature Selection for Text Classification: A Case Study on Phishing Email Detection,” International Journal of Information Engineering and Electronic Business, vol. 7, no. 2, p. 60, 2015.
7. N. Cheng, R. Chandramouli, and K. Subbalakshmi, “Author gender identification from text,” Digital Investigation, vol. 8, no. 1, pp. 78–88, 2011.
8. J. Verbeek, “Supervised feature extraction for text categorization,” in Tenth Belgian-Dutch Conference on Machine Learning (Benelearn’00), 2000.
9. S. L. Lam and D. L. Lee, “Feature reduction for neural network based text categorization,” in Database Systems for Advanced Applications, 1999. Proceedings., 6th International Conference on, pp. 195–202, IEEE, 1999.
10. A. Selamat and S. Omatu, “Web page feature selection and classification using neural networks,” Information Sciences, vol. 158, pp. 69–88, 2004.

11. J.-T. Sun, Z. Chen, H.-J. Zeng, Y.-C. Lu, C.-Y. Shi, and W.-Y. Ma, "Supervised latent semantic indexing for document categorization," in Data Mining, 2004. ICDM'04. Fourth IEEE International Conference on, pp. 535–538, IEEE, 2004.
12. N. Slonim and N. Tishby, "Document clustering using word clusters via the information bottleneck method," in Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval, pp. 208–215, ACM, 2000.
13. J. C. Gomez, E. Boiy, and M.-F. Moens, "Highly discriminative statistical features for email classification,"
14. Knowledge and information systems, vol. 31, no. 1, pp. 23–53, 2012.
15. S. Karamizadeh, S. M. Abdullah, A. A. Manaf, M. Zamani, and A. Hooman, "An overview of principal component analysis," Journal of Signal and Information Processing, vol. 4, no. 03, pp. 173–175, 2013.
16. M. Lichman, "UCI machine learning repository," 2013. [16] G. Forman and E. Kirshenbaum, "Extremely fast text feature extraction for classification and indexing," in Proceedings of the 17th ACM conference on Information and knowledge management, pp. 1221–1230, ACM, 2008.
17. H. Uguz, " ˇ A two-stage feature selection method for text categorization by using information gain, principal component analysis and genetic algorithm," Knowledge-Based Systems, vol. 24, no. 7, pp. 1024–1032, 2011.
18. H. Uguz, " ˇ A multistage feature selection model for document classification using information gain and rough set," International Journal of Advanced Research in Artificial Intelligence(IJARAI), vol. 3, no. 11, 2014.
19. S. Vidhya, D. A. A. G. Singh, and E. J. Leavline, "Feature Extraction for Document Classification," International Journal of Innovative Research in Science,Engineering and Technology(IJIRSET), vol. 4, no. 6, pp. 50– 56, 2015.
20. M. F. Porter, "An algorithm for suffix stripping," Program, vol. 14, no. 3, pp. 130–137, 1980.
21. M. F. Porter, "Effective PreProcessing Activities in Text Mining using Improved Porter's Stemming Algorithm," International Journal of Advanced Research in Computer and Communication Engineering, vol. 2, no. 12, pp. 4536–4538, 2013.
22. C. Cortes and V. Vapnik, "Support-vector networks," Machine learning, vol. 20, no. 3, pp. 273–297, 1995. [23] Z. Li, R. Yuan, and X. Guan, "Accurate classification of the internet traffic based on the svm method," in Communications, 2007. ICC'07.IEEE International Conference on, pp. 1373–1378, IEEE, 2007.
23. C.-C. Chang and C.-J. Lin, "LIBSVM: a library for support vector machines," ACM Transactions on Intelligent Systems and Technology (TIST), vol. 2, no. 3, p. 27, 2011.
24. J. Wang and X. Li, "An improved KNN algorithm for text classification," in Information Networking and Automation (ICINA), 2010 International Conference on, vol. 2, pp. V2–436, IEEE, 2010.
25. G. Guo, H. Wang, D. Bell, Y. Bi, and K. Greer, "KNN modelbased approach in classification," in OTM Confederated International Conferences" On the Move to Meaningful Internet Systems", pp. 986– 996, Springer, 2003.
26. M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The weka data mining software: an update," ACM SIGKDD explorations newsletter, vol. 11, no. 1, pp. 10–18, 2009.

AUTHORS PROFILE

A. Ravi Kumar has completed B.E from karnatak university, Dharwad. M.Tech from Bharath University, Chennai and pursuing PhD from SriSatya Sai University of Technology & Medical Sciences, Sehore, Bhopal. He has 13 Yrs of teaching Experience. His interesting areas are Software Engineering, Cloud computing and Bigdata. Has published more than 20 papers in various journals and conferences.

Dr. G. Anil Kumar currently working as the principal in Scient Institute of Technology, Hyderabad. Has graduated from Osmania university. He obtained his masters in software engineering from Jawaharlal Technological University Hyderabad, Telangana State. He was awarded Doctorate from the JNTU Hyderabad. Has guided a number of B.Tech and M.Tech (CSE) students for their projects. He has attended several national and international conferences. He delivered lectures in various Engineering colleges in and around Hyderabad. He published over 25 research papers in various National and International journals/ conference proceedings. Has 16 years of teaching experience and 1 year of industry experience.