TEXT CATEGORIZATION USING SUPPORT VECTOR MACHINE CLASSIFICATION

[1]Ms.P.Ranjitha, [2]Mr.N.Sunilprasanth, [3]Mr.M.Dineshkumar
[1]Assistant Professor, [2]Student, [3]Student
[1]Department of computer science and engineering,
[1]Sri Krishna College Of Technology, Tamil Nadu, India

Abstract : Text categorization is a vital and all around considered territory of example acknowledgment, with an assortment of present day applications. Viable spam email sifting frameworks, mechanized report association and administration, and enhanced data recovery frameworks all advantage from systems inside this field. The issue of highlight determination or picking the most applicable highlights out of what can be an extraordinarily huge arrangement of information, is especially essential for exact Text categorization. The proposed framework (I) utilize understood pre-handling technique doorman and Lancaster for prepare the dataset. (ii) various component choice measurements have been investigated in content arrangement, among which data pick up (IG), chi-square (CHI), Mutual data (MI), Ng-Goh-Low (NGL), Galavotti-Sebastiani-Simi (GSS), Relevancy Score (RS), Multi-Sets of Features (MSF) Document recurrence (DF) and chances proportions (OR) are viewed as best. Pruning strategies are additionally proposed utilizing overlook the component in view of TF and DF to additionally decrease the arrangement of conceivable highlights (normally words) inside a report preceding applying a technique for include determination. (iii) Finally order the chose include in light of two calculation bolster vector machines and Navie bayes. Two benchmark accumulations were picked as the testbeds: Reuters-21578 and little bit of Reuters Corpus Version 1 (RCV1). The two classifiers and the two information accumulations, and that a further increment in execution is acquire by consolidating uncorrelated and high-performing highlight choice techniques

IndexTerms **– Reuters-21578, Reuters Corpus Version 1, chi-squars, etc...**

I. INTRODUCTION

Content classification is a vital and all around considered territory of example acknowledgment, with an assortment of present day applications. Viable spam email sifting frameworks, mechanized report association and administration, and enhanced data recovery frameworks all advantage from systems inside this field. The issue of highlight determination or picking the most applicable highlights out of what can be an extraordinarily huge arrangement of information, is especially essential for exact content classification. The proposed framework (I) utilize understood pre-handling technique doorman and Lancaster for prepare the dataset. (ii) various component choice measurements have been investigated in content arrangement, among which data pick up (IG), chi-square (CHI), Mutual data (MI), Ng-Goh-Low (NGL), Galavotti-Sebastiani-Simi (GSS), Relevancy Score (RS), Multi-Sets of Features (MSF) Document recurrence (DF) and chances proportions (OR) are viewed as best. Pruning strategies are additionally proposed utilizing overlook the component in view of TF and DF to additionally decrease the arrangement of conceivable highlights (normally words) inside a report preceding applying a technique for include determination. (iii) Finally order the chose include in light of two calculation bolster vector machines and Navie bayes. Two benchmark accumulations were picked as the testbeds: Reuters-21578 and little bit of Reuters Corpus Version 1 (RCV1). The two classifiers and the two information accumulations, and that a further increment in execution is acquire by consolidating uncorrelated and high-performing highlight choice techniques.

II. TEXT CATEGORIZATION

Classification of information mining frameworks as per the sort of information sources mined: This arrangement is as indicated by the kind of information took care of, for example, spatial information, interactive media information, time-arrangement information, content information, world wide web, and so forth.

Classification of data mining systems according to the database involved: This order in view of the information demonstrate included, for example, social database, protest arranged database, information distribution center, value-based database, and so on.

Classification of data mining systems according to the kind of knowledge discovered: This arrangement in view of the sort of learning found or information mining functionalities, for example, portrayal, separation, affiliation, characterization, bunching,

and so on. A few frameworks have a tendency to be far reaching frameworks offering a few information mining functionalities together.

Classification of data mining systems according to mining techniques used: This order is as per the information investigation approach utilized, for example, machine learning, neural systems, hereditary calculations, insights, perception, database arranged or information distribution center situated, and so forth.

III. PROBLEM DEFINITION

Data recovery and content mining techniques work on the terms found in content archives. All things considered, each term found in a gathering is investigated and utilized for additionally preparing. The procedure of highlight choice is performed keeping in mind the end goal to lessen the quantity of terms to be utilized as a part of further investigation (i.e. to recognize the most imperative terms heretofore). The undertaking of this venture is to look at a scope of highlight determination methods with the objective of a careful execution assessment.

IV. IMPLEMENTATION

4.1 Pre-Processing

4.1.1 Removal of stop words

In the greater part of the applications, it is viable to expel words which show up time and again (in each or relatively every archive) and in this way bolster no data for the assignment. Great cases for this sort of words are relational words, articles and verbs like" be" and" go". In the event that the crate" Apply stop word evacuation" is checked, every one of the words in the document" swl.txt" are considered as stop words and won't be stacked. This document contains as of now the 100 most utilized words in the English dialect which by and large record for a half of all perusing in English. On the off chance that the container" Apply stop word evacuation" is unchecked, the stop word expulsion calculation will be incapacitated when the corpus is stacked.

4.1.2 Stemming

Stemming or lemmatisation is a procedure for the diminishment of words into their root. Numerous words in the english dialect can be lessened to their base frame or stem e.g. concurred, concurring, deviate, assention and contradiction have a place with concur. Besides, are names changed into the stem by evacuating the" 's". The variety" Peter's" in a sentence is lessened to" Peter" amid the stemming procedure. The consequence of the evacuation may prompt an off base root. Be that as it may, these stems don't need to be an issue for the stemming procedure, if these words are not utilized for human connection. The stem is as yet helpful, on the grounds that every single other intonation of the root are changed into a similar stem.

4.2 Feature selection

4.2.1 odds ratio

Chances Ratio thinks about the chances of a component happening in one classification with the chances for it happening in another classification. It gives a positive score to highlights that happen more frequently in one class than in the other, and a negative score on the off chance that it happens more in the other. A score of zero means the chances for an element to happen in one class is precisely the same as the chances for it to happen in the other, since ln (1) = 0. The first Odds Ratio calculation for twofold order:

$$OP(\Phi, X_\kappa) = \lambda\nu \frac{P(F|C_k)(1-P(F|\overline{C_k}))}{P(F|\overline{C_k})(1-P(F|C_k))} = \lambda\nu \frac{\left(\frac{N_{F,C_k}}{NC_k}\right)\left(1-\frac{N_{F,\overline{C_k}}}{N\overline{C_k}}\right)}{\left(\frac{N_{F,\overline{C_k}}}{N\overline{C_k}}\right)\left(1-\frac{N_{F,C_k}}{NC_k}\right)} \quad (1)$$

$$\Pi(\Phi|X_\kappa) = \frac{N_{F,C_k}}{N_{C_k}} \quad (2)$$

Let P (t|c) be the likelihood of an arbitrarily picked word being t, given that the report it was looked over has a place with a class c. At that point chances (t|c) is characterized as P (t|c)/[1– P (t|c)] and the Odds Ratio equivalents to,

$$OP(\tau) = \lambda\nu\,[o\delta\delta\sigma(\tau|\chi+) / o\delta\delta\sigma(\tau|\chi\square)] \quad (3)$$

Clearly, this scoring measure favors includes that are illustrative of positive cases. Thus, a component that happens not very many circumstances in positive reports yet never in negative archives will get a generally high score

4.2.2 Information gain

Here both class enrollment and the nearness/nonappearance of a specific term are viewed as irregular factors, and one figures how much data about the class participation is picked up by knowing the nearness/nonattendance measurements (as is utilized as a part of choice tree enlistment. Undoubtedly, if the class enrollment is translated as an arbitrary variable C with two

esteems, positive and negative, and a word is similarly observed as an irregular variable T with two esteems, present and missing, at that point utilizing the data theoretic meaning of shared data we may characterize Information Gain as:

I.

$$I\Gamma(\tau) = H(X) - H(X|T) = \sum_{\chi} \sum_{\tau} \Pi(X=\chi, T=\tau) \log[\Pi(X=\chi, T=\tau)/\Pi(X=\chi)\Pi(T=\tau)] \quad (4)$$

Here, τ goes over {present, absent} and c extends over {c+, c– }. As pointed out over, this is the measure of data about C (the class mark) picked up by knowing T (the nearness or nonappearance of a given word).

4.2.2 Document frequency (DF) Thresholding

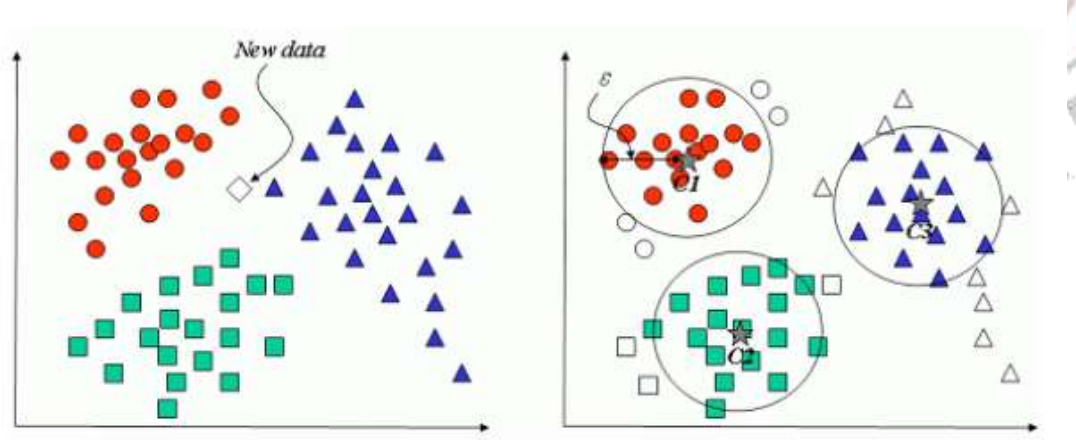
One of the easiest techniques for vocabulary diminishment, and thus vector dimensionality lessening, is the Document Frequency Thresholding

$$\Delta\Phi(\Phi) = N_{\Phi} \quad (5)$$

The quantity of records containing a component in the preparation set is checked. This is improved the situation each component in the preparation set, before evacuating all highlights with a report recurrence not as much as some predefined edge and highlights with a recurrence higher than some other limit. Then again, the archive recurrence can be utilized as some other element determination technique where it makes a positioned rundown, and returns the most astounding positioned highlights.

4.3 Text Classification

An outstanding content arrangement technique is KNN other prominent strategies are Naive Bayesian proposed the Centroid based classifier and demonstrated that it gives preferred outcomes over other known techniques. Demonstrate that expelling exceptions from the preparation classes altogether enhances the order comes about got with KNN strategy. Our examinations demonstrate that the new technique gives preferable outcomes over the Centroid-based classifier. Before applying our grouping calculation, we will expect that a TFxIDF record by-term network has been made and disguised and that the P-tree adaptation of the lattice has additionally been made. We will just work on the P-tree variant. To classify another record, d new, the calculation first endeavors to discover the k-most-comparative neighbors.



The study comprised of non-financial companies listed at KSE-100 Index and 30 actively traded companies are selected on the bases of market capitalization.And 2015 is taken as base year for KSE-100 index.

5.1 Navie Bayesian Text Classification Algoritham

The Naïve Bayesian arrangement framework depends on Bayes' control and functions as takes after. There are classes, say Ck for the information to be arranged into. Each class has a likelihood P(Ck) that speaks to the earlier likelihood of grouping a property into Ck; the estimations of P(Ck) can be assessed from the preparation dataset. For n trait esteems, vj, the objective of order is plainly to locate the restrictive likelihood P(Ck | v1 ∧ v2 ∧ ⋯ ∧ vn). By Bayes' lead, this likelihood is equal to

$$\frac{P(v_1 \wedge v_2 \wedge \ldots \wedge v_n \mid C_k) P(C_k)}{P(v_1 \wedge v_2 \wedge \ldots \wedge v_n)}$$

For grouping, the denominator is immaterial, since, for given estimations of the vj, it is the same paying little heed to the estimation of Ck. The focal suspicion of Naïve Bayesian grouping is that, inside each class, the qualities vj are for the most part free of each other. At that point by the laws of free likelihood, for grouping, the denominator is insignificant, since, for given estimations of the vj, it is the same paying little mind to the estimation of Ck. The focal supposition of Naïve Bayesian arrangement is that, inside each class, the qualities vj are on the whole autonomous of each other. At that point by the laws of free likelihood,

P(vi | {all alternate estimations of vj}, Ck) = P(vi | Ck) and in this manner

P(v1 ∧ v2 ∧ … ∧ vn | Ck) = P(v1 | Ck)P(v2 | Ck)… P(vn | Ck).

Each factor on the right-hand side of this condition can be resolved from the preparation information, in light of the fact that (for a discretionary vi),

P(vi | Ck) ≈ [#(vi ∧ Ck)]/[#(Ck)]

where "#" speaks to the quantity of such events in the preparation set information. In this manner, the grouping of the test set would now be able to be assessed by,

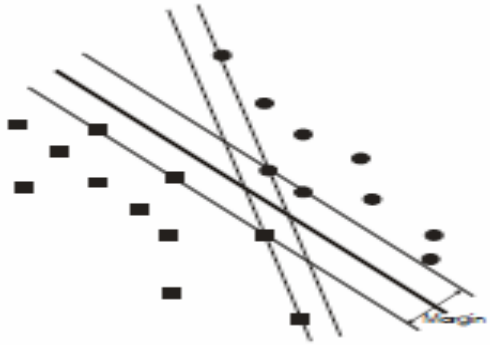
P(Ck | v1 ∧ v2 ∧ … ∧ vn) which is relative to

P(Ck) P(v1 | Ck) P(v2 | Ck) P(v3 | Ck) … P(vn | Ck).

As specified over, the focal presumption in Naïve Bayesian order is that given a specific class enrollment, the probabilities of specific characteristics having specific esteems are free of each other.

5.2 Support Vector Machine

SVMs are a for the most part appropriate apparatus for machine learning. Assume we are given with preparing cases xi, and the objective esteems y_i (- 1,+1)¬. SVM scans for an isolating hyperplane, which isolates positive and negative cases from each other with maximal edge, at the end of the day, the separation of the choice surface and the nearest case is maximal



The equation of a hyperplane is:

$$w^T x + b = 0 \quad (6)$$

The classification of an unseen test example **x** is based on the sign of $w^T x + b$. The separator property can be formalized as:

$$w^T x_i + b \geq 1 \; if \; y_i = +1$$
$$w^T x_i + b \leq 1 \; if \; y_i = -1$$

The optimization problem of SVM is the following: Minimize over (**w**,b) the value of $\frac{1}{2}.w^T w$ subject to:

$$\forall_{i=1}^{n}: y_i[w^T x_i + b] \geq 1 \quad (7)$$

Systematic risk is the only independent variable for the CAPM and inflation, interest rate, oil prices and exchange rate are the independent variables for APT model.

$$\frac{1}{2}.w^T w + C.\sum_{i=1}^{n} \varepsilon_i \quad (8)$$

subject to slack variables: minimize over (w,b) the value of

$$\forall_{i=1}^{n}: y_i[w^T x_i + b] \geq 1 - \varepsilon_i \quad (9)$$

Where C is a steady to tradeoff amongst edge and preparing error.Second, with piece capacities. Let K(x_i,y_j) indicate a capacity that generally gives how comparative two cases are. This is known as a bit work on the off chance that it fulfills the Mercers' condition. The least difficult case is when K(x_i,y_j)=x_i^T.x_j, however more entangled cases makes SVM appropriate for non-straightly distinguishable issues.

An intriguing property of SVM is that the ordinary of the choice surface is a straight mix of cases. This implies, the choice capacity can be

$$f(x) = sign\left(\sum_{i=1}^{N} \alpha_i . y_i . K(x_i, x) + b\right) \quad (10)$$

Machine learning algorithms tend to overlearn when the dimensionality is high, for example when more dimension exists than example.

V. Conclusion

In the choice procedure, each component (term or single word) is allocated with a score as indicated by a score-registering capacity. At that point those with higher scores are chosen. These numerical meanings of the score-registering capacities are regularly characterized by a few probabilities which are assessed by some measurement data in the reports crosswise over various classes. Content grouping is a directed procedure that utilizations marked preparing information to take in the characterization framework and afterward naturally orders the rest of the content utilizing the scholarly framework. Characterization assumes a crucial part in numerous data administration and recovery assignments. In such manner, we first endeavor to apply some content pre-process in various dataset, and after that we separate an element vector for each new report by utilizing highlight weighting and highlight choice calculations for upgrading the content grouping exactness. In Experiments, albeit the two calculations demonstrate adequate outcomes for content grouping,

References

[1] W. Lam, M. Ruiz, and P. Srinivasan, "Automatic text categorization and its application to text retrieval," IEEE Trans. Knowl. Data Eng., vol. 11, no. 6, pp. 865–879, Nov./Dec. 1999.

[2] G. Forman, "An extensive empirical study of feature selection metrics for text classification," The J. Mach. Learn. Res., vol. 3, pp. 1289–1305, 2003.

[3] H. Liu and L. Yu, "Toward integrating feature selection algorithms for classification and clustering," IEEE Trans. Knowl. Data Eng., vol. 17, no. 4, pp. 491–502, Apr. 2005.

[4] P. M. Baggenstoss, "The pdf projection theorem and the class-specific method," IEEE Trans. Signal Process., vol. 51, no. 3, pp. 672–685, Mar. 2003.

[5] A. McCallum and K. Nigam, "A comparison of event models for naive Bayes text classification," in Proc. Workshop Learn. for Text Categorization, 1998, vol. 752, pp. 41–48.

[6] D. Koller and M. Sahami, "Hierarchically classifying documents using very few words," in Proc. 14th Int. Conf. Mach. Learn., 1997, pp. 170–178.

[7] Y. H. Li and A. K. Jain, "Classification of text documents," The Comput. J., vol. 41, no. 8, pp. 537–546, 1998.

[8] T. Joachims, "Text categorization with support vector machines: Learning with many relevant features," in Proc. 10th Eur. Conf. Mach. Learn., 1998, pp. 137–142.

[9] S. Eyheramendy, D. D. Lewis, and D. Madigan, "On the naive Bayes model for text categorization," in Proc. 9th Int. Workshop Artif. Intell. Statist., 2003, pp. 332–339.

[10] L. Galavotti, F. Sebastiani, and M. Simi, "Experiments on the use of feature selection and negative evidence in automated text categorization," in Proc. 4th Eur. Conf. Res. Adv. Technol. Digit. Libraries, 2000, pp. 59–68.

[11] B. Tang, S. Kay, and H. He, "Toward optimal feature selection in naive Bayes for text categorization," Preprint, arXiv:1602.02850 [stat.ML], 2016.

[12] L. Wang, N. Zhou, and F. Chu, "A general wrapper approach to selection of class-dependent features," IEEE Trans. Neural Netw., vol. 19, no. 7, pp. 1267–1278, Jul. 2008.

[13] B. Tang, H. He, Q. Ding, and S. Kay, "A parametric classification rule based on the exponentially embedded family," IEEE Trans. Neural Netw. Learn. Syst., vol. 26, no. 2, pp. 367–377, Feb. 2015.
[14] D. Cai, Q. Mei, J. Han, and C. Zhai, "Modeling hidden topics on document manifold," in Proc. 17th ACM Conf. Inf. Knowl. Manage., 2008, pp. 911–920.
[15] B. Tang and H. He, "KernelADASYN: Kernel based adaptive synthetic data generation for imbalanced learning," in Proc. IEEE Congress Evol. Comput., 2015, pp. 664–671.