TEXT CATEGORIZATION: BY COMBINING NAÏVE BAYES AND MODIFIED MAXIMUM ENTROPY CLASSIFIERS

[1]AADITYA JAIN, [2]R. D. MISHRA

[1] M.Tech Scholar, [2] Processor,
Department of Computer Science & Engg., R. N. Modi Engg. College,
Rajasthan Technical University, Kota, Rajasthan, India
E-mail: [1]aadityajain58@gmail.com, [2]rdmishracs@gmail.com

Abstract— Text Categorization is done mainly through classifiers proposed over the years, Naïve Bayes and Maximum Entropy being the most popular of all. However, the individual classifiers show limited applicability according to their respective domains and scopes. Recent research works evaluated that the combination of classifiers when used for categorization showed better performance than the individual ones. This work introduces a modified Maximum Entropy based classifier. Maximum Entropy classifiers provide a great deal of flexibility for parameter definitions and follow assumptions closer to real world scenario. This classifier is then combined with a Naïve Bayes classifier. Naïve Bayes Classification is a very simple and fast technique. The assumption model is opposite to that of Maximum Entropy. The combination of classifiers is done through operators that linearly combine the results of two classifiers to predict class of documents in query. Proper validation of the 7 proposed modifications (4 modifications of Maximum Entropy, 3 combined classifiers) are demonstrated through implementation and experimenting on real life datasets.

Keywords— Text Categorization, Combining Classifiers, Naïve Bayes Classifier, Maximum Entropy Classifier

I. INTRODUCTION

The amount of text available for analysis has increased hugely in recent years due to the social networking, micro blogging and various messaging/ bulletin board systems. Besides these, many articles, news feeds and documents are now available in soft copy. An important step in text classification is to classify the text documents among some known set of classes/ categories.

The task of data mining can be done through two processes- Classification and Clustering. While clustering is an unsupervised learning approach, classification is a supervised form of machine learning. It helps to classify the given text in different categories using efficient classification algorithms. The classification process in itself is a very detailed process consisting of various stages. Each stage then has a set of methods to choose from depending on the text and the given classification problem. The final stage is the classification stage. Algorithms, called Classifiers, are trained using documents already classified (manually) and then used to predict the category of a new text document. Over the years, many classification algorithms have been proposed, out of which Naïve Bayes [1], Maximum Entropy [2], K-Nearest Neighbor (KNN) [3] and Support Vector Machines (SVM) [4] are commonly used till now. Each classifier is restricted to its scope of classification which makes it difficult to effectively classify the given text. Extensions of the classifiers were also proposed to overcome the drawbacks.

Yigit et al. [5] used KNN classifier for detecting news related to Turkey among the different news channels. The classification process carried out by the KNN classifier was found to be 90% accurate. Similarly, Naïve Bayes outperforms SVMs for Authorship attribution in [6]. Such research works bring us to the conclusion that each classifier works well only on specific applications. Therefore, each upcoming application will have to be tested against various available classifiers to find which classifier works well. A generalized classification algorithm is therefore needed which suits to every application. Hybridization of classifiers in order to bring about the best of combined classifiers is found to be a promising approach in this direction. The results of combinations when compared to the results of the individual classifiers are visibly better which give a boost to this area of research.

In this paper, Naïve Bayes [1] and Maximum Entropy classifiers [2] are considered for combination for the purpose of text classification. Whereas Naïve Bayes is extremely simple, the Maximum Entropy classifier provides great flexibility and uniformity. The assumption models of both differ completely. Naïve Bayes assumes total independence between words in the document (which is realistically impossible) unlike Maximum Entropy classifier which is approximate to the real world scenarios. Modifications to the traditional Maximum Entropy classifier are proposed making it more efficient and then the modified versions of Maximum Entropy Classifier are combined with the Naïve Bayes classifier using three merging operators-Max, Average and Harmonic Mean. The performance is measured on different datasets such that no individual classifier has clearly better performance over all of them.

II. LITRATURE SEARCH

This section reviews some relevant hybrid approaches for the purpose of text classification. Recent research

works in the direction of combining classifiers for text classification assure that combination is always better than using individual classifiers.

As early as in 1996, Larkey and Croft [7] propose the combination of three classifiers, KNN, Relevance feedback and Bayesian's independence classifiers to be used in the medical domain for automatic assignment of ICD9 codes. The task was done first with individual classifiers and then with combined to check the effectiveness of both the approaches and the hybrid approach was concluded better. The performance of the classifiers were measured based on document ranks. This is an example where classifiers are used for document ranking. The approach is of using weighted linear combination.

Bennett, Dumais and Horovitz [8] proposed a probabilistic method to combine the classifiers such that the contribution of a classifier depends on its reliability. The reliability is measured through reliability indicators which are linked to the regions where a classifier might perform relatively good or poor. Instead of the rank of document, the indicators are based on performance of the classifier itself thus making the proposal more generalized.

Grilheres, Brunessaux and Leray [9] published detailed study of effect of combining classifiers to classify multimedia documents into heterogeneous classes. Various combinations are applied to a five thousand web pages document of the European Research Project Net Protect II and experiment results prove that with a prior knowledge on classifiers, better filtering performances are possible. The approaches used for combining are both voting-based and logic-based.

Besides the conventional style of linear or voting based combination a new technique based on Dampster-Shafer theory was proposed by Sarinapakkornand Kubat [10]. Their main aim is fusion of sub-classifiers since the application is towards multi-label classification.

Isa et al in their two successive papers [11] and [12] have proposed a novel idea as to how meta-outputs of a Naïve Bayes technique can be used with SVM and Self-organizing maps (SOM) respectively. Bayes formula is used to convert the text document into a vector space where the values denote the probabilities of documents towards any class depending on the features contained. This is called the vectorisation phase of the classifier. It is common to both the classifiers. SVM is then applied on this vector space model for final classification output. The proposal had improved classification accuracy compared to the pure naive Bayes classification approach. In [12] the probability distributions obtained by Bayes technique are followed by an indexing step done through SOM to retrieve the best match cases. SOM is similar to clustering of documents based on a similarity measure between the documents like Euclidean distance.

Miao et al [13] considered very different combination of classifiers, namely KNN and Rocchio methods. A variable precision rough set is used to partition the feature space to lower and upper bounds of each class. Each subspace is classified through Rocchio technique. But it fails when the arriving document is in boundary region, here kNN is used. This presents a new style of combining classification methods to overcome each others' drawbacks.

Fragos, Belsis and Skourlas [14] also concludes in favor of combining different approaches for text classification. The methods that authors have combined belong to same paradigm – probabilistic. Naïve Bayes and Maximum entropy classifiers are chosen to test on the applications where the individual performance is good. The merging operators are used above the individual results. Maximum and Harmonic mean operators have been used and the performance of combination is better than the individual classifiers.

Keretna, Lim and Creighton [15] have worked on recognizing named entities from a medical dataset containing informal and unstructured text. For this, they combine the individual results of Conditional Random Field (CRF) classifiers and Maximum Entropy (ME) classifiers on the medical text; each classifier trained using a different set of features. CRF concentrates on the contextual features and ME concentrates on the linguistic features of each word. The combined results were better than the individual results of both the classifiers based on Recall rate performance measure.

Ramasundaram [16] aimed to improve the N-grams classification algorithmby applying Simulated Annealing (SA) search technique to the classifier. The hybrid classifier NGramsSA brought about an improvisation to the original NGrams classifier while inheriting all the advantages of N-grams approach. Feature reduction using χ^2 method is used but its multivariate value among the n-grams affects the performance of the classifier.

III. PROPOSED CLASSIFIER

This section discusses the document model used for representing the text documents, the modified classifiers considered for the combination and the proposed classification process.

3.1. Representation of Document Model

For representing documents, term frequency matrix is used which tells the number of times a particular term has appeared in the document. Each document is M-tuple of values, where each value is frequency of the term occurring in the document D_i, that is $d_i = \langle t_{i1}, t_{i2}, \dots, t_{iM} \rangle$ as shown in the following matrix

$$\begin{bmatrix} t_{11}t_{12}t_{13} \dots\dots\dots & t_{1M} \\ t_{21}t_{22}t_{23} \dots\dots\dots & t_{2M} \\ \vdots & \vdots \\ t_{N1}t_{N2}t_{N3} \dots\dots\dots & t_{NM} \end{bmatrix}$$

The notations to be taken care of here are discussed below.

- C represents the number of classes;
- M represents the number of features/ terms;
- R represents the number of documents in the training set;
- S represents the number of documents in the testing set;
- N represents the total number of documents; that is $N = R + S$;

3.2. Naïve Bayes Classifier

The Naive Bayes classifier [1] is considered one of the simplest of probabilistic models showing how the data is generated with the following assumption

"Given the context of the class, all attributes of the text are independent to each other."

This technique starts by taking text documents as word counts. It calculates the class conditional probability followed by the classification probability or posterior probability to be used by the trained classifier to predict the class of any document.

For every term t_i and class c_j, the class conditional probability $\hat{P}(t_i|c_j)$considering only one training set is given as follows:

$$\hat{P}(t_i|c_j) = \frac{1 + number\ of\ times\ t_i\ appears\ in\ a\ document\ from\ class\ c_j}{d + number\ of\ words\ in\ all\ documents\ from\ class\ c_j}$$

$$\hat{P}(t_i|c_j) = \frac{\sum tf(t_i, d \in c_j) + \alpha}{\sum N_{d \in c_j} + \alpha . M} \quad (1)$$

Where,

$\sum tf(t_i, d \in c_j)$:The sum of raw term frequencies of word t_i from all documents in the training sample that belong to class c_j.

α: An additive smoothing parameter

$\sum N_{d \in c_j}$: The sum of all term frequencies in the training dataset for class c_j .

The posterior probability of a document belonging to any class c_j. is the product of individual class-conditional probabilities of all terms contained in the query document.

$$P(d|c_j) = \hat{P}(t_1|c_j).\hat{P}(t_2|c_j) \dots\dots \hat{P}(t_M|c_j)$$
$$= \prod_{i=1}^{M} \hat{P}(t_i|c_j)^{tf(t_i,d)} \quad (2)$$

Once all these probabilities have been computed, the maximum probability towards a class c_k indicates that query document d belongs to class c_k.

$$k = \text{argmax}_j P(d/c_j) \quad (3)$$

3.3. Maximum Entropy Classifier

Maximum Entropy classifier [2] believes in the principle that the model generating the training set should be the most uniform among the other models and all constraints from the training set should be satisfied in the model.

Let, $f(d, c)$ be the feature function of the document with the class;

$p(c|d)$be the required probability that assigns class c to document d;

and$\tilde{p}(c|d)$ be the empirical probability distribution;

Then, maximum entropy principle says that expected value of $f(d, c)$ is same for both $p(c|d)$ and $\tilde{p}(c|d)$.This can be called a constraint which makes

$$p(c|d) = \frac{1}{Z(d)} \exp\left[\sum_i \lambda_i f_i(d, c)\right] \quad (4)$$

Here,

$Z(d) = \sum_c \exp[\sum_i \lambda_i, f_i(d, c)]$is normalization factor and λ_i is weight for each feature $f_i(d, c)$

3.4. Modified Maximum Entropy Classifier

The ME Classifier is modified in three aspects weightsλ_i, featuresf_i and Prediction Probability.

The weights λ_i can be computed using any of the weighting methods Gini Index, Chi square, CMFS or DIA instead of the conventional method of optimizing the objective function in ME Classifier. These weighting methods are discussed below:

- Gini Index

Suppose S_i is the sample set which belongs to class c_j, s is the sample number of setS_i, then the Gini index of set S is:

$$Gini(S) = 1 - \sum_{i=1}^{n} (P(S_i/s))^2 \quad (6)$$

- Chi-Square (CHI)

Chi-square formula is defined as follows:

$$CHI(t_k, c_i) = \frac{N(a_{ki}d_{ki} - b_{ki}\, c_{ki})^2}{(a_{ki} + b_{ki})(a_{ki} + c_{ki}) + (b_{ki} + d_{ki})(c_{ki} + d_{ki})} \quad (7)$$

Where$N$ is the amount of documents in the training set; a_{ki} is the frequency with which feature t_k occurs in the category c_i ; b_{ki} is the frequency with which feature t_k occurred in all categories except c_i; c_{ki} is the frequency with which categoryc_i occurs and does not contain featuret_k; $d_{ki}$is the number of times neither c_i nor t_k occurs.

- DIA Association Factor (DIA)

The DIA association factor is defined by

$$DIA(t_k, c_i) = P(c_i|t_k) \quad (8)$$

where $P(c_i|t_k)$refers to the conditional probability that feature t_k belongs to category c_i when the feature t_k occurs.

- Comprehensive Measurement Feature Selection (CMFS)

$$CMFS(t_k, c_i) = P(t_k|c_i)P(c_i|t_k) \quad (9)$$

- Global Feature Selection (GFS)

This feature selection algorithm is the global version of CMFS and involves sum of CMFS for all classes c_i thereby improving the performance of the classification. It is defined by

$$FS(t_k) = \sum_{c_i} P(t_k|c_i)P(c_i|t_k) \quad (10)$$

Features f_i are computed as feature contribution towards a class using

$$f_i = f(t_i, c_j) = \frac{Sum\ of\ term\ frequencies\ of\ t_i\ for\ class\ c_j}{Total\ number\ of\ terms\ of\ class\ c_j} \quad (11)$$

The prediction probability of ME classifier has been modified as

$$P(d|c_j) = \frac{exp[\sum_t \lambda(t_i, c_j) f(t_i, c_j) tf(d, t_i)]}{\sum_j exp[\sum_t \lambda(t_i, c_j) f(t_i, c_j) tf(d, t_i)]} \quad (12)$$

Our proposed modification involves multiplication of weights (λ_i) and feature function (f_i) with term frequency (tf) for calculation of prediction probability unlike the conventional method used for ME Classifier.

3.5. Proposed Combined Classifiers

Stage by stage representation of the classification process is illustrated in Fig.1.

The proposed classification process consists of the following stages

- Preprocessing stage
- Feature Extraction stage
- Individual Classification stage
- Combining Classification stage
- Final Results

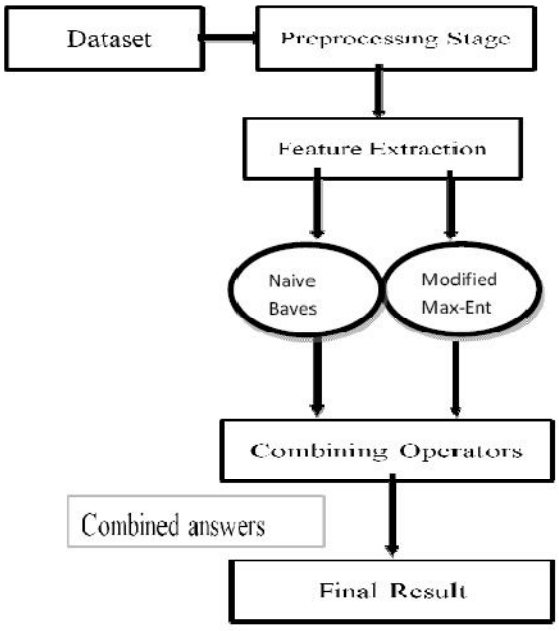

Fig. 1: Classification Process

The classification process starts with preprocessing of text to make it ready for the classification followed by extracting relevant features through Global Feature selection (GFS) method. It ranks the features according to their importance and the top K features are extracted. After the feature extraction process, Naïve Bayes (NB) and Maximum Entropy (ME) classifiers are used individually for classification. The later stage combines both the classifiers using three combination operators: Average, Harmonic Mean and Max. Combining operators are used for compensation of errors in each classifier and performance improvement. Equations (13), (14) and (15) show the Average, Max and Harmonic Mean operators respectively.

$$Average(d) = \text{avg}(NB(d), ME(d)) \quad (13)$$

$$Max(d) = \max(NB(d), ME(d)) \quad (14)$$

$$Harmonic(d) = \frac{(2.0 * NB(d) * ME(d))}{(NB(d) + ME(d))} \quad (15)$$

The results of the combination then give the final result.

3.6. Classification Algorithm

- Input: Training data DR as term frequencies and class labels and test document d, number of classes C.
- Output: Predict class C for document d.
- Step 1: Train NB Classifier's class conditional probability using tf and class labels as per Eqn(1).
- Step 2: Compute posterior class probability of NB, P_{NB} for every class as per Eqn (2).
- Step 3: Train ME Classifier: Compute λ_i using any of the weighing schemes CHI-Square, DIA factor, Gini Index or CMFS and feature function f_i as in Eqn (11).
- Step 4: Compute posterior probability using ME, P_{ME} for every class using Eqn (12).
- Step 5: Combining Results: Compute for every class, probability that d belongs to c_j as
  $P(d|c_j) = \text{COMB}\left(P_{ME}(d|c_j), P_{NB}(d|c_j)\right)$
  Where COMB represents any one of the operator either Max, Harmonic Mean or Average.
- Step 6: Predict class of d as class for which $P(d|c_j)$ is maximum by Eqn (3).

Among the Naïve Bayes and Maximum Entropy Classifiers tested individually, Naïve Bayes Classifier gives the best results in most of the cases in spite of its assumption of total independence of words in the document which does not actually happen in real world scenarios. The modifications in ME are

proposed for overcoming the drawbacks of the original ME classifier related to high training time complexity. So we can say that the proposed combination classifier using Max operator that is Combo Classifier MNBME gives the best results.

CONCLUSIONS

In this paper, combination of classifiers with some major modifications has been done. Naïve Bayes is combined with modified Maximum Entropy classifiers; Naïve Bayes for its simplicity and Maximum Entropy classifier for its flexibility and appropriateness to the real world scenarios. Both the classifiers are opposite with respect to the assumption model; the former is a totally independent model while the latter considers entropy relation among terms. The modified versions of Maximum Entropy classifiers have the original Maximum Entropy classifiers with new methods for the computation of weights, feature functions and prediction probability. The task of splitting datasets is done by distributing a specified percentage of documents in the training set with the remaining documents in the test set. The ratio of distribution may or may not vary for different datasets. The modified versions of Maximum Entropy classifier are combined with Naïve Bayes using any of the Average, Max or Harmonic Mean operators.

The datasets for experiments have been selected such that in few cases Naïve Bayes performs better than Modified Maximum Entropy classifier and the opposite in few; while for others both classifiers have equivalent performance. Given any such case, the proposed combination classifier with Max combining operator gives the best accuracy.

REFERENCES

[1] D. Lewis, "Naive Bayes at Forty: The Independence Assumption", Information Retrieval. Proc. ECML-98, 10th European Conf. Machine 1998.

[2] K. Nigam, J. Lafferty, and A. McCulllum, "Using Maximum Entropy for Text Classification", IJCAI-99, Workshop on Machine learning for Information Filtering, pgs 61-67, 1999.

[3] G. Guo, H. Wang, D. Bell, Y. Bi, and K. Greer, "KNN Model-Based Approach in Classification", Proc. ODBASE pp- 986 – 996, 2003.

[4] C. Basu and M. S. Waters, "Support Vector Machines for Text Categorization", Proc. 36th Annual Hawaii International Conference on System Sciences, 2003.

[5] FerruhYiğit, ÖmerKaanBayka, "Detection of The News about Turkey with Web-based Text Mining System", International Journal of Information Technology & Computer Science (IJITCS, Volume 11, Issue No: 2, pp.56-6-, 2013.

[6] FatmaHowedi and MasnizahMohd, "Text Classification for Authorship Attribution Using Naive Bayes Classifier with Limited Training Data", Computer Engineering and Intelligent Systems, Vol.5, No.4, 2014.

[7] L.S. Larkey. and W. B. Croft, "Combining classifiers in text categorization",Proc. SIGIR-96, 19th ACM International Conference on Research and Development in Information Retrieval (Zurich, CH, 1996), pp. 289–297 1996.

[8] Paul N. Bennett, Susan T. Dumais, Eric Horvitz, "Probabilistic Combination of Text Classifiers Using Reliability Indicators: Models and Results", Proceedings of 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Tampere, Finland, August 2002. ACM Press.

[9] B. Grilheres, S. Brunessaux, and P. Leray, "Combining classifiers for harmful document filtering", RIAO '04 Coupling approaches, coupling media and coupling languages for information retrieval, Pages 173-185, 2004.

[10] KanoksriSarinnapakorn and MiroslavKubat, "Combining Subclassifiers in Text Categorization: A DST-Based Solution and a Case Study", IEEE Transactions On Knowledge And Data Engineering, Vol. 19, No. 12, December 2007.

[11] Dino Isa, Lam Hong lee, V. P Kallimani, and R. Raj Kumar, "Text Documents Preprocessing with the Bayes Formula for Classification using the Support vector machine", IEEE Transactions of Knowledge and Data Engineering, vol.20, no. 9, pp.1264-1272, September 2008.

[12] Dino Isa, V. P Kallimani and Lam Hong lee, "Using Self Organizing Map for Clustering of Text Documents", Expert System with Applications, vol. 36, no. 5, pp. 9584-9591, July, 2009.

[13] Duoqian Miao , QiguoDuan, Hongyun Zhang, and Na Jiao, "Rough set based hybrid algorithm for text classification", Journal of Expert Systems with Applications, vol. 36, no. 5, pp. 9168-9174, July 2009.

[14] K. Fragos, P.Belsis, and C. Skourlas, "Combining Probabilistic Classifiers for Text Classification",Procedia - Social and Behavioral Sciences, Volume 147 Pages 307–312, 3rd International Conference on Integrated Information(IC-ININFO), doi: 10.1016 /j.sbspro .2014.07. 098 , 2014.

[15] S. Keretna, C. P. Lim and D. Creighton, "Classification Ensemble to Improve Medical Named Entity Recognition", 2014 IEEE International Conference on Systems, Man, and Cybernetics, San Diego, CA, USA, 2014.

[16] S.Ramasundaram, "NGramsSA Algorithm for Text Categorization", International Journal of Information Technology & Computer Science (IJITCS), Volume 13, Issue No : 1, pp.36-44, 2014.

★★★