Question Classification using Machine Learning Approaches

Arun D Panicker
Dept of Computer Science
Amrita School of Engineering

Athira U
Dept of Computer Science
Amrita School of Engineering

Sreesha Venkitakrishnan
Dept of Computer Science
Amrita School of Engineering

ABSTRACT
Question classification is the process by which a system analyzes a question and labels the question based on the category to which it belongs. The automated categorization (or classification) of questions into predefined categories has witnessed a booming interest due to the increased popularity of web technologies. The recent advancement in the form of E-Learning calls for the need of question categorization. In network based learning the questions posted by students need to be categorized on the basis of the concerned concepts. This point to the relevance of question categorization in this area. Many approaches to question classification have been proposed and have achieved reasonable results. The dominant approaches are machine learning and context based classification. There are several Machine Learning methods for question categorization. Here we are extending the previous methods for text categorization to question categorization and making a comparative study of the performance of two approaches, Naïve Bayes and Support Vector Machine

General Terms
Machine Learning Algorithms

Keywords
Question Categorization; Naïve Bayes; SVM; Entropy

1. INTRODUCTION
With the emergence of question answering systems the rate of posting and answering questions by people increased. As a result of which system accumulated large number of questions. Hence, it is necessary to organize these questions in a good way. Question categorization is a technique used for this purpose. Question Categorization, is a useful technique in Web-based Question Answering system. On the basis of the questions, it will be associated to the corresponding category.

Earlier approaches for the creation of automatic document classifiers consisted of manually building, by means of *knowledge engineering* (KE) techniques, an expert system capable of taking Document Categorization decisions. [8],the major disadvantage of which was that it required rules manually defined by a knowledge engineer with the aid of a domain expert. Another problem is that when the classifier is ported to a completely different domain, concerned domain expert need to intervene and the work has to be done in its entirety.

To overcome the pitfalls associated with rule-based classification 'Machine Learning' techniques are currently applied for these purposes. In this approach set of pre-classified questions are fed to the classifier. This acts as the training example for the classifier. Based on these examples the classifier will classify the future samples. The common text classifiers which employ these approaches include probabilistic classifiers, decision tree classifiers, decision rule classifiers, regression based classifiers, neural network based classifiers, and SVM based classifiers [2].

Another approach that can be taken is context based interpretation [3]. It takes advantage of tracking the contextual meaning of words and phrases during (and after) the development of ontology for that context, and subsequently uses this information as knowledge base for interpretation of free text sentences.

Here we are proposing two approaches, SVM and Naïve Bayes, which have been previously used for text classification, for network-based learning wherein the questions posted by students on-line will be classified into their corresponding categories thus reducing the task of the tutor in finding out questions related to one particular portion. The proposed method brings about variation in ordinary text classification by incorporating a different weight calculation method to account for question categorization. The classifiers will be trained by set of training examples for each category, which are predefined. Hence forth the classifiers will be used to classify set of questions. The performance evaluation of classifiers using both approaches, SVM and Naïve Bayes, is conducted.

The paper is organized as follows. Section 2 describes about Naïve Bayes Classifier and the algorithm which has been employed for question categorization using this approach. Section 3 describes about Support Vector Machine and the algorithm employed for question categorization. Section 4 describes the evaluation. Section 5 gives a comparative study of the approaches and section 6 refers to the conclusion and future works.

2. NAÏVE BAYES CLASSIFIER
A Naive Bayes classifier is a simple probabilistic classifier based on applying Bayes' theorem with strong independence assumptions [7]. Bayes theorem can be stated as follows

$$P(c_k / q_j) = \frac{P(c_k) * P(q_j / c_k)}{P(q_j)} \qquad (1)$$

Where $P(C_k|q_j)$ is the posterior probability,

$P(C_k)$ is the prior probability,

$P(q_j| C_k)$ is the likelihood and $P(q_j)$ is the evidence

A naive Bayes classifier follows conditional independence since it assumes that the presence (or absence) of a particular feature of a class is unrelated to the presence (or absence) of any other feature, given the class variable. Thus terms are given a weight value which is independent of its position and presence of other terms. Naive Bayes classifier is trained by

set of labeled training examples. Thus it is said to undergo supervised learning.

2.1 Algorithm

Naive Bayes algorithm proposed by us proceeds by finding out the feature vectors associated with each category. Feature vector includes common terms occurring in questions pertaining to one particular category expressed in terms of their weight or relevance in particular question. Common terms are found out by stop-word elimination, stemming and then pruning (eliminating words with frequency below a particular range and frequency above a particular range).Weight of a term in a category, associated with a particular question can be found out using weight calculation methods. Here the weight is calculated by finding the entropy associated with the term [9]. This is given by

$$a_i = 1 + \frac{1}{\log(N)} \sum \frac{f_{it}}{n_i} \log \frac{f_{it}}{n_i} \qquad (2)$$

Where a_i is the weight associated with word$_i$,
N denotes total number of categories,
fit : frequency of word$_i$ in question t
n_i : total number of occurrences of word *i* in all questions

The next phase of algorithm is to classify a new question. The probability of the question to belong to all categories are found out and the category for which it has maximum posterior probability is the one to which the question is assigned to.

$$c_k = argmax_{c_k} P(c_k / q_j) \qquad (3)$$

where C_k is the category with maximum posterior probability.

The probability of a question q_j to belong to a category c_k is given by (1)

Since Naive Bayes classifier assumes conditional independence P(q_j | c_k) can be given as

$$P(q_j / c_k) = \prod P(a_i / c_k) \qquad (4)$$

$$P(a_i / c_k) = p_i^{a_i} (1 - p_i)^{1 - a_i} \qquad (5)$$

The weight calculation followed here is different from ordinary tf-idf method followed for text classification. The classification using the above said approach of weight calculation showed increased accuracy. The accuracy was found to increase with increase in number of training data.

To speed up the calculation of weight the weight and terms have been saved in database. MySQL database has been used. Index has been provided which further improved the ease of retrieval thus reducing the classification time.

3. SUPPORT VECTOR MACHINE

A Support Vector Machine model is a representation of the examples as points in space, mapped so that the examples of the separate categories are divided by a clear gap that is as wide as possible. New examples are then mapped into that same space and predicted to belong to a category based on which side of the gap they fall on. a support vector machine constructs a hyperplane or set of hyperplanes in a high or infinite-dimensional space, which can be used for classification, regression, or other tasks[2]. Intuitively, a good separation is achieved by the hyperplane that has the largest distance to the nearest training data points of any class (so-called functional margin), since in general the larger the margin the lower the generalization error of the classifier.

3.1 Optimal Separating Hyperplane

Consider a two classes and use labels −1/+1 for the two classes. The sample is X={xt , rt} where

r t = +1 if **xt** ∈ C1 and

rt = −1 if **xt** ∈ C2.

We would like to find w and w0 such that

(rt *(*w ^{T} ***xt*** + *w0* *)* ≥ +1

The distance from the hyperplane to the instances closest to it margin on either side is called the margin, which we want to maximize for best generalization. Now that we are using the hypothesis class of lines, the optimal separating hyperplane is the one that maximizes the margin [2].

The distance of **xt** to the discriminant is

$$\frac{|w^T x^t + w_0|}{\|w\|}$$

which, when rt ∈ {−1,+1}, can be written as

$$r^t \frac{(w^T x^t + w_0)}{\|w\|} \qquad (6)$$

and we would like this to be at least some value ρ.

We would like to maximize ρ but there are an infinite number of solutions that we can get by scaling w and for a unique solution, we fix ρ||w||= 1 and thus, to maximize the margin, we minimize ||w||.

In our proposal we proceed by converting questions into its numerical representation. This representation is obtained by representing each question by its probability to belong to a category. Thus each question will have as many dimensions as the number of categories. Each dimension having value which is equal to the probability of the question to belong to that category. .Thus for n categories each question will have n dimension and ith dimension will have the value which is equal to the probability of the question to belong to the category i.

3.2 Algorithm

1. Read the training data from the input file.

2. Remove stop words

3. Find the probability of each question

4. For each pair of classes (0 to N*(N-1)/ 2), find the discriminant function associated with the pair (C1, C2).

5. Use α to define the hyperplane that discriminates the instances of the two classes C1 & C2.

6. The weight vector associated with the hyperplane is given by:

$$Wi = \sum \alpha i \, Si \qquad (7)$$

7. Equation of the separating hyperplane is given by:

y = wx + w0, where w0 is the bias. (8)

8. Read the questions to be classified and classify it on the basis of this discriminant function defined as:

If ((y = wx + w0) > 1)

Data lies on the positive side of the hyperplane i.e, it belongs to the class C1 (increment the vote of C1)

Else if ((y = wx + w0) < -1)

Data lies on the negative side of the hyperplane i.e. it belongs to the class C2.

Else

Data lies on the margin.

9. The test data belong to the class with maximum vote.

argmaxi (vote(Ci))

4. EXPERIMENTS

Experiments were conducted based on two approaches namely Naïve Bayes and SVM for question categorization. We have conducted experiments on a training set comprising of 1500 questions, each for twenty categories of twenty newsgroup. The questions pertaining to these categories have been prepared from the newsgroup categories.

In the case of Naïve-Bayes average of 1120 keywords were identified for all categories. The weights for the words were assigned as per the equation(5).The probability of these words to belong to corresponding classes were also found. When a new question is encountered the keywords in it were identified and the weight of the keyword and its probability to belong a particular class is found. The probability for the question to belong to a class is found by using the equation (2). This is repeated for all the classes. The class for which this value is highest is the one to which the question belongs.

For example the question "What is Darwin fish?" is to be classified.

The keywords identified and their corresponding weights are as follows

Darwin: 0.0148457

Fish: 0.0147090

On following the above mentioned steps the probability for the question to belong to the category 1 has been found as: 0.919153, which was the highest. Thus the question has been classified to category 1 i.e.alt.atheism.

In the case of SVM average of 1120 keywords were identified for all categories. Weights associated with each keyword of a question to belong to particular class have been found out using equation (7). This is repeated for all classes. A question can be represented as a set of features whose number is equivalent to the number of categories. The ith feature is the root of squared sum of the weights of the keywords in that question to belong to the ith category. The weight matrix associated with the questions is found. For a new question, the keywords have been identified and features are calculated

5. CONCLUSION

We have proposed methods for question classification using Naïve Bayes and SVM. The Naive Bayes Classifier is a very popular algorithm due to its simplicity, computational efficiency and its surprisingly good performance for real-world problems. But it is not capable of solving more complex classification problems. The experimental results show that SVM consistently achieve good performance on question categorization task, outperforming the Naïve Bayes approach in these cases. From our experimental results we can conclude that SVM is very promising and easy to use method in question classification. Hence SVM based classifier can be suggested as an effective classifier for E-Learning applications. The above classifier can be extended to accommodate network based technologies so that the classifier can be used to automatically classify the questions posted by student over the world, on line. This relieves the tutors from the task of sorting the questions based on the topic, as the system automatically classifies the question and drops it under concerned category.

This will serve as the input to the SVM and corresponding y-value is calculated. Depending on the value of y the question will be classified to the appropriate category.

Weight vector for the training set has been found out. Feature vector for the test question mentioned above has been obtained as

[210 0]

The y-value corresponding to the feature vector is -3.4994 which is a negative value .Thus the question is classified to category1.

For classification tasks, the terms true positives, true negatives, false positives, and false negatives compare the results of the classifier under test with trusted external judgments. The terms positive and negative refer to the classifier's prediction (sometimes known as the observation), and the terms true and false refer to whether that prediction corresponds to the external judgment (sometimes known as the expectation).

The performance evaluation of Naïve Bayes and SVM has been done in terms of precision and recall

TABLE1. Precision for categories alt.atheism and comp. Graphics using the classifiers Naïve- Bayes and SVM

Classifier	Category1	Category2
NaïveBayes	1	0.58
SVM	0.95	0.95

TABLE2. Recall for categories alt.atheism and comp. Graphics using the classifiers Naïve- Bayes and SVM

Classifier	Category1	Category2
NaïveBayes	0.75	0.92
SVM	0.95	0.95

The precision and the recall show that Naïve Bayes classifier is found to be an effective classifier for most of the data. But as the complexity of the data increases Support Vector Machine forms an effective classifier. This is because the process of finding the probability associated with each word in a new question is a cumbersome task as the number of questions increase. Thus Naïve Bayes proves to be less efficient. Further Support Vector Machine uses overfitting protection. SVMs have the potential to handle large feature spaces. SVM takes into account only few irrelevant features

Another possible future work is to enable its application in the area of E-Governance where ordinary people can post their queries and submit it without insisting them to specify the area of query as the classifier automatically classifies the query to concerned topic. This relieves ordinary people from the technical knowledge to move the query to concerned category.

The approaches followed here do not consider the word disambiguity i.e. the classification of question containing words that can belong to different categories for e.g. word "data" can belong to class data base as well as data structure. The work can be extended to include probability assignment of such words and finding out correct class to which it can be associated.

6. ACKNOWLEDGMENTS

We would like to record our profound gratitude to Dr.Ramachandra Kaimal of Computer Science Department, Amrita school of Engineering for his motivation and direction towards the preparation of this paper. We would also like to express our gratitude to Computer science Department, Amrita School of Engineering for providing us with facilities to complete our project.

7. REFERENCES

[1] HAYES, P. J., ANDERSEN, P. M., NIRENBURG, I. B., AND SCHMANDT, L. M. Tcs: shell for content-based text categorization. In Proceedings of CAIA-90, 6th IEEE Conference on Artificial Intelligence Applications (Santa Barbara, CA, 1990), 320–326., 1990

[2] ETHEM,ALPAYDIN,Introduction to Machine Learning, MIT Press, Cambridge.,2010.

[3] XIAOSHAN,PAN., AND FRANZ ,J. KURFESS.A Context-Based Free Text Interpreter,California Polytechnic State University San Luis Obispo Master's Thesis - Computer Science Department

[4] S.DUMAIS.,Improving the retrieval information from external sources. Behavior Research Methods, Instruments and Computers, 23:229–236., 1991.

[5] KOLLER, D., AND SAHAMI, M.Hierarchically classifying documents using very few words. In Proceedings of ICML-97, 14th International Conference on Machine Learning (Nashville, TN, 1997), 170–178.,1997.

[6] J. KIVINEN, M. WARMUTH., and P. AUER. The perceptron algorithm vs. winnow: Linear vs. logarithmic mistake bounds when few input variables are relevant. In Conference on Computational Learning Theory., 1995.

[7] T. MITCHELL. Machine Learning. McGraw-Hill, New York,NY., 1997.

[8] SEBASTIANI,F. Machine Learning in Automated Text Categorization, Consiglio Nazionale delle Ricerche, Italy.,ACM Computing Surveys.,2002.

[9] HUANG,P.,BU,J.J., CHEN,C., AND QIU, GUANG.An Effective Feature-Weighting Model for Question Classification,International Conference on Computational Intelligence and Security.,2007