Using Supervised Clustering Technique to Classify Received Messages in 137 Call Center of Tehran City Council

Mahdiyeh Haghiri[1 *], Hamid Hassanpour[2]
(1) Information Technology engineering/e-commerce, Shiraz University
(2) Department of Computer Eng. & IT, Shahrood University of Technology,Shahrood,Iran
Mahdiyehaghiri.66@gmail.com; h_hassanpour@yahoo.com

Received: 2011/05/20; Accepted: 2011/06/14 Pages: 15-24

Abstract

Supervised clustering is a data mining technique that assigns a set of data to predefined classes by analyzing dataset attributes. It is considered as an important technique for information retrieval, management, and mining in information systems. Since customer satisfaction is the main goal of organizations in modern society, to meet the requirements, 137 call center of Tehran city council is planning to reduce the waiting time of customers, forwarding their messages to the appropriate service manager to increase service quality. The city council is currently using a manual approach dispatching textual request messages. Since this process is very similar to supervised clustering concept, in this study, we applied the Naïve Bayes algorithm, which is one of the most common algorithms in the supervised clustering arena to classify received messages regarding their subjects. The performance results of the proposed technique indicate its high efficiency in clustering the received messages with 98% accuracy.

Keywords: *supervised clustering, classification, Naïve Bayes algorithm, data mining*

1. Introduction

Nowadays, due to the growing development of communication and information technologies, our society needs to take advantages of human and social resources as much as possible. In this regard, for instance, reducing the waiting time of customers, increasing service quality and simplifying the communications are needed in social communications in this age.

Because of the growing development of IT-based communication tools and easy access to telecommunications systems such as phones, mobiles, computers and the internet, and due to engagement in doing routine tasks, using simple and available technologies has a special importance. Therefore, Tehran municipality on a new plan made a system by using information technology science and telecommunication tools. So, because of the direct view and citizens' active collaboration and people's involvement in managing their own environment, civil tasks can be done quickly and properly.

137 call center of Tehran is founded as an infrastructure in order to facilitate communications of citizens. The received messages are classified according to their content into predefined subject categories in the system by human operators and referred to the proper unit to be surveyed more.

Because of the implementation process in 137 call center that is naturally based on specific and predefined choices, this paper presents a model to classify the received messages automatically. The proposed model uses a well known technique for textual document clustering as an important step in indexing, retrieval, management and mining of data in information systems [1]. Clustering methods are classified into two groups: supervised and unsupervised clustering.

Supervised clustering also known as classification [2], is a data mining technique used for extracting model from data on discrete values [3]. Classification evaluates the features of a data set, and then assigns them labels from a predefined set. This is the most common capability of data mining.

Data classification is a two-phase process [3]. In the first phase, that is called learning step, classification system is made by a predefined set of data labels or concepts. In this phase, the classification algorithm makes the classifier by analyzing or learning from training data set.

In the second phase - generalization step [4], trained model is used to classify unlabeled data. A test set is used to estimate the classifier performance. The main goal of generalization phase is to decrease noise impact on the classification results and increase classifier accuracy. Classifier accuracy on a given test set is the percentage of tuples that have been correctly classified. If classifier accuracy is admitted, it can be used to classify tuples that their class labels are unknown.

Naïve Bayes, Rocchio, K-Nearest Neighbors (K-NN), Regression, Decision Trees, Neural Networks, Support Vector Machines (SVM) and Decision Rule, can be referred to as the most important methods for text classifications problem [5].

Unsupervised clustering – or simply clustering [2], is in fact an unsupervised learning method. The goal is to divide and gather a set of unstructured objects into appropriate clusters or groups so that similar objects appear in the same cluster, and different objects in the separate clusters [6,7].

The remainder sections of this paper are as follows. Section 2 explains our approach in detail. Following, Section 3 introduces our dataset and some related challenges related to Persian language of our dataset. Section 4 is devoted to experimental results, and finally section 5 offers conclusions and suggests future work.

2. Introducing the proposed algorithm

The classification algorithm is used to classify received messages from citizens in the 137 call center of Tehran city council based on Naïve Bayes algorithm. To know more about the mechanism of this kind of algorithm, we will explain it in great details in this section.

Assume we have m classes $C_1, C_2, \ldots, C_m$. Naïve Bayes classification algorithm assigns X record to a class, say C_i, which has the maximum posterior probability:

$$P(C_i \mid X) > P(C_j \mid X) \quad for \quad 1 \le j \le m \quad , \quad i \ne j \tag{1}$$

With this assumption, $P(C_i|X)$ maximizes the posterior probability.

$$P(C_i \mid X) \frac{P(X \mid C_i) P(C_i)}{P(X)} \tag{2}$$

Since the value of P(X) is the same for all classes, to maximize (2), $P(X|C_i)P(C_i)$, it is enough to maximize $P(C_i|X)$. Additionally if $P(C_i)$, that is named prior probability is equal for all classes, it is only required for $P(X|C_i)$ to be maximized. If the prior probability of classes is different, it is possible to compute $P(C_i)$ from (3):

$$P(C_i)\frac{|PC_i, D|}{|D|} \tag{3}$$

where $|C_i, D|$ is the number of records of class C_i in the dataset D, and $|D|$ is the total number of records in the data set.

The amount of $P(X|C_i)$ computed by (4):

$$P(X\ |C_i) = \Pi_{k=1}^{n} P(X_k\ |C_i) \tag{4}$$

In this relation, X_k is the amount of attribute of A_k for record X and n is the total number of all attributes.

If attributes are discrete, then:

$$P(X\ |C_i) = \frac{P(X_k\ |C_i)}{|C_i, D|} \tag{5}$$

Since our calculation for classifying a record X, is based on the percentage of occurring keywords of special class C_i (weight of the keyword for a subject), formula (6) is used to compute the weight of each keyword for each subject:

$$P\left(X\ |C_i\right) = \frac{thefrequencyofeachkeywordforsubject C_i}{thefrequencyofeachkeywordforallsubjects} \tag{6}$$

a. Offered architecture for classification

Offered architecture in this paper can classify the received messages from citizens to predefined subjects. This algorithm is presented in figure (1), weights keywords of each message for every subject and then calculates the weight of the subjects for each message by summing the weights of the keywords in each message. Finally, it assigns to the message the subject with maximum weight.

```
Algorithm SAMANE (Input Message: String, Output: Classification of Message)
Begin
   1- Select a Message;
   2- Give-Token(Message, Token-Set);
   3- Find New Tokens in Relation by Tokens in Token-Set and Insert Token-Set and Unite with
      Related Token;
   4- For Each Token in Token-Set Do
      Begin
      Find All Message in Database in Which It Occurs and Insert It in Message-Set;
      4-1- For Each Message in Message-Set Do
      Begin
      Find All Themes (Subjects) Related to Message;
      End
      Themes-Frequency=Frequency of All Themes for Each Token;
      4-2- For Each Theme in Theme-Set Do
      Begin
      Weight (Token, Theme) = Frequency of Themes (Subjects)/ Themes-Frequency;
      End
   End
   5- For Theme=1 to Number of Themes Do
      Begin
      For Token=1 to Number of Tokens Do
         Begin
         Sum (Theme) = Sum(Theme)+Weight(Token, Theme);
         End
   End
   6- Max=Sum (1);
     For Theme=2 to Number of Themes Do
     Begin
     If Sum(Theme)>Max Then Max=Sum(Theme);
     End
   7- Return Class=Subject(Max);
End.
```

Figure 1. offered classification algorithm

Table (1) represents a sample of the amount of the keywords relation with predefined subjects for each message. This assumed message has n keywords that have been already existed in m subjects of all predefined subjects.

Table 1. The relation between keywords and defined subjects

Subject / Word	Subject m	……	Subject 2	Subject 1
Keyword 1	15%	……	40%	27%
Keyword 2	20%	……	25%	10%
……	……	……	……	……
Keyword n	5%	……	30%	36%

This algorithm estimates the weight of each keyword in existing m subjects and calculates the weight of each message for each probable subject by computing the total assigned weights of the keywords in each subject. The way of calculating the weight of each keyword in different subjects has been shown in (6).

The message classification steps are as following:

1. *Select a message from received messages;*
2. *Find the keywords existing in the message;*
3. *Find synonyms and words relating to the keywords and integrate them;*
4. *Make Token-Set (set of the keywords and the synonyms relating to the message);*
5. *Send Token-Set to the weight calculation unit for each subject;*
6. *Calculate the weights for each subject for message;*
7. *Assign the message to the subject with the maximum weight;*
8. *Register information so that it can be used for searching and retrieving.*

b. The components of the offered architecture

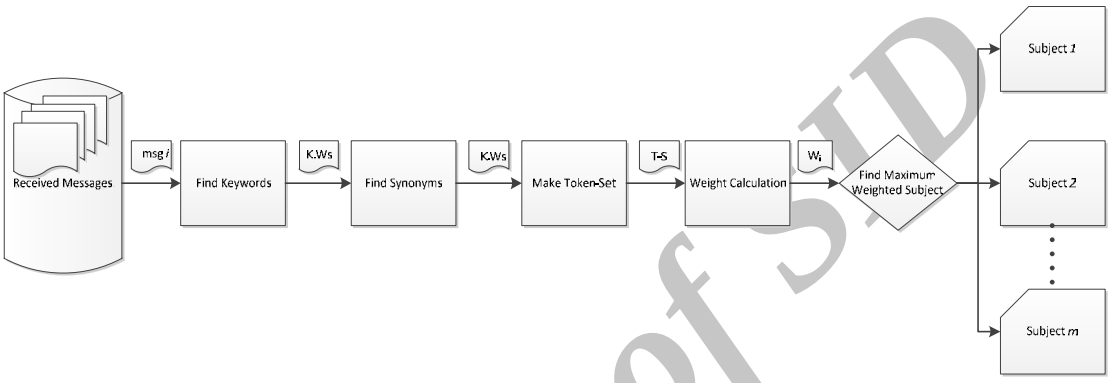

Figure 2. offered architecture for classification

i. Training data set

Training data are sets of information that transfer a kind of knowledge to the learner agent as a presupposed one. Actually, basic difference between two methods of clustering results from the difference between their learning methods. Training information given to the system has been discussed in the following.

- *Class Labels*

The 137 call center of Tehran city council has classified the received messages from citizens into 44 specific classes and then refers them to the relevant unit. Since we plan to classify automatic data based on the current process, and also we want to provide the same condition to compare the two methods, 44 classes with the same labels are included in the offered system. Moreover a class called "other cases" has been considered so that a received message out of the knowledge extent of the system can be classified.

- *Training messages*

Each of the 44 classes in the manual system contains some related messages. These messages have been chosen and discussed by a group of human expertise assigned by Tehran municipality and include all types of probable problems in every subject. These messages have been known as training messages in our system and introduced to the system by their class labels.

- *Keywords*

Preprocessing explores simple data in order to change them to appropriate format for training [8]. One of the preprocessing phases is to extract keywords. Each document has one or more keywords or key phrases describing its content, where these keywords or key phrases belong to a finite set called controlled dictionary [5].

If there is a set of training documents with distinctive keywords for each of them, the process of extracting keywords will be a supervised learning, otherwise will be an unsupervised learning [8].

Due to the nature of our data training, we always face the short messages. Our target keywords are also exactly technical in the urban management and belong to the predefined classes. So identifying and defining these few technical words which many of them have been already known increases the accuracy and efficiency of the system, and requires less time regarding to identifying all the useless words such as additional words, conjunctions, pronouns, adverbs, plural suffixes, verbs and, etc. In this paper we profit from supervised learning for keyword extraction.

The offered system has been designed in a way that it can distinguish defined keywords from all the suffixes, prefixes, punctuations such as ".", ",", ";", "?", etc. and also multi syllabus words. So the keywords have been introduced to the system in their simplest faces.

ii. Test data set

The data which have been used to test the offered system, are the sent messages from the citizens in the first five months in 2011in the districts 1,2, 5 and 7 of the municipality of the region 1 in Tehran. These 356 messages that have been already classified manually are in 20 classes out of 44 training classes. In this paper, it is supposed that citizens' messages should be sent in Persian.

c. Overcoming Persian language challenges

In [9], some of the Persian language challenges in computer especially in the internet and result of web searching, have been mentioned. These challenges are as following:

1. *Variety in using clung or unclung «می» such as «می‌تواند» and «میتواند»;*
2. *Variety in using clung or unclung «ها» such as «درخت‌ها» and «درختها»;*
3. *Using some prefixes or suffixes such as «پاک‌سازی» and «پاکسازی»;*
4. *Using «همزه» in various types such as «مسئول» and «مسؤول» or «مسأله» and «مسئله»;*
5. *Using or not using «ۀ» such as «خانۀ مسکونی» and «خانه مسکونی»;*
6. *Variety in using «ی» in Arabic words ended by «ا» such as «موسی» and «موسا»;*
7. *Variety in spelling some words such as «اتاق» and «اطاق»;*
8. *Using European words in native language or Persian translations such as «ATM» and «خودپرداز»;*
9. *Using or not using irregular plurals for some words;*
10. *Transforming Latin words to Persian alphabets with the original pronunciation such as «source» and «سورس»;*
11. *Using «ا» and «آ» interchangeably such as «قرآن» and «قران»;*
12. *Using or not using phonetic symbols for words.*

The only difference between our data and what found in computer, and the internet is that our data are messages sent by their own cell phones or mobiles. Due to the present abilities of mobiles in our country, cases such as 5 and 12 can't be set forth for discussion. On the other hand, citizens are supposed to send their messages in Persian, so cases 8 and 10 can be ignored, too.

Because of the system capacity to extract keywords from all of possible suffixes and prefixes, cases 1 and 2 have been corrected.

Our offered system has a complete controlled dictionary that keywords belong to it. This dictionary in addition to synonymous words and phrases of keywords, consists of different spelling of them so that the system can respond to the cases 3, 6, 7 and 9 correctly.

But for cases 4 and 11, an ability is considered in the system so that all the writing styles of vowels such as «و» ,«ا» and «ی» can be changed to the native writing styles and then it will survey them.

3. Evaluation the offered system performance

To estimate algorithm performance, three scales Precision(P), Recall(R) and Accuracy(A) were surveyed[10]. Four parameters TP, FN, FP and TN[11] that are defined in table (3), are needed to calculate these scales.

Table 3. parameters of efficiency scales

Text	Assigning to class C_i	Not assigning to class C_i
Belonging to class C_i	True Positive(tp)	False Negative(fn)
Belonging to non class C_i	False Positive(fp)	True Negative(tn)

Accuracy is the common scale used to evaluate accuracy of machine learning algorithms. This scale is calculated by (7)[12]:

$$A = \frac{tp + tn}{tp + tn + fp + fn} \tag{7}$$

Precision and Recall are scales used to estimate efficiency of the text classification algorithm. Precision scale calculates the accuracy of a searching. This scale is calculated by (8)[12]:

$$P = \frac{tp}{tp + fp} \tag{8}$$

Recall scale, unlike the precision scale, calculates the rate of perfection in a searching [12]:

$$R = \frac{tp}{tp + fn} \tag{9}$$

Table (4) presents the efficiency scales of the system in all target classes. The average of this asset value is used for final evaluation.

Table 4. Performance metrics related to some target classes

Class name	Accuracy	Precision	Recall
water	0.977	1	0.91
bus	0.991	0.75	0.86
asphalt	0.983	1	0.84
Social damages	0.997	0.8	1
snow	0.985	0.83	0.88
Parks and green lands	0.985	0.66	0.79
traffic	1	1	1
Changing usage	0.997	1	0.88
Removing and pitching	0.958	0.91	0.8
animals	0.992	0.73	1
trees	0.983	0.89	0.94
trashes	0.966	0.71	0.77
building	0.983	0.95	0.9
Blocking path	0.98	1	0.53
washing	1	1	1
dredge	0.977	1	0.75
repairing	0.989	0.89	0.89
trouble	0.952	0.64	0.43
Pitching safety marks	0.989	0.67	0.67
cleaning	0.98	1	0.61
Average	0.984	0.87	0.82

The results from the surveys showed that the system has totally made acceptable errors in the following conditions:

1. *The nearness of the message content to two or more groups with the high usage and meaning resemblance;*
2. *The presence of different usage and meaning keywords in the message;*
3. *Introducing the problem in a statement, without giving any solution or requesting to solve the problem.*

The first error resulting from the high meaning resemblance and usage occurs because the right range of the subject has not been paid attention.

For words such as park that has different meanings in different sentences, it is necessary to extract not only the defined keywords but also survey the other parts of the sentence so that the target meaning in each sentence can be specified through the relevant words with each possible meaning.

The third error occurs because the range of some groups is too wide. Although the citizens can help survey their message better and faster by requesting the municipality in their own messages, at least by making their own requests besides the introduced problems.

Therefore, to correct the possible errors and increase the efficiency of the system as much as possible, the following actions can be taken.

1. *Reforming the existing groups by merging or splitting the groups of less accuracy;*

2. *Using a method besides using a dictionary to index latent semantic in the sentences made up of the different applicable meaning keywords;*
3. *Appropriate teaching and informing the citizens to ask their own requests in the message.*

However, despite the introduced problems the rate of offered system precision is 87%, and its Recall is 82%. Of course, we should notice that these acceptable percentages have considerably improved and makes the system efficiency increase by overcoming the presented difficulties. The accuracy measure calculated is over 98%. This shows one of the most important advantages of the Naïve Bayes algorithm in comparison with other algorithms that is its persistence against the irrelevant attributes and keywords.

4. Conclusion

In this paper, a method based on one of the most common supervised algorithms, Naïve Bayes, was presented to classify the citizens' thematic messages in the 137 call center of Tehran city council. This algorithm uses 44 existing classes in the manual system as training classes, 605 reference messages as its training messages and 356 received messages from Tehran citizens in district 4 of Tehran municipality of region 1 for five months in a year, as its test data.

Applying this method and evaluating its results proved this algorithm with accuracy up to over 98%, is a right choice for our target. The rate of precision 87% and recalling 82% in classifying the messages also suggests the comparative success of the offered system in classifying the target test data. Moreover, more study on the obtained results showed that the efficiency of the system can be increased by reforming the target groups used in the 137 call center and subsequently in the system and adding more convenience to the capacity of indexing its latent semantic, and teaching and informing the citizens correctly in presenting their problems and requests. Our suggested future work consists of using other methods beside the current algorithm to increase accuracy.

5. References

[1] Azzag, H., Guinot, C. and Venturini, G., "Data and text mining with hierarchical clustering ants" Springer, Swarm Intelligence in Data Mining. Vol. 34,: 153-189, 2006.

[2] Eick, C., F., Zeidat, N. and Zhao, Z., "Supervised Clustering- Algorithms and Benefits" 16th IEEE International Conference on Tools with Artificial Intelligence (ICTAI'04): 774-776, 2004.

[3] Mazhari, N., Imani, M., Joudaki, M. and Ghelichpour, A., " An overview of classification and its algorithms" 3th Data Mining Conference (IDMC'09): Tehran, 2009.

[4] Fazaelijavan, M. and Sadredini, M., H., " Investigating Classification and Association Rule Mining Techniques for Intrusion Detection" 3th Data Mining Conference (IDMC'09): Tehran, 2009.

[5] Sebastiani, F., "Machine learning in automated text categorization" Journal of ACM Computing Surveys., Vol. 34, 2002.

[6] Finley, T. and Joachims, T., "Supervised clustering with support vector machines" ICML '05 Proceedings of the 22nd international conference on Machine learning, 2005.

[7] Berkhin, P., "Survey Of Clustering Data Mining Techniques" Springer. Vol. 12: 1-56, 2002.

[8] Ahmadi, M., H., Monadjemi, A., H. and Ayat, S., "Persian Text Classification Using Association Rules" 4th Data Mining Conference (IDMC'10): Tehran, 2010.

[9] محسنی‌کبیر، نینا، مینایی‌بیدگلی، بهروز و کاشفی، امید، "مطالعه و بررسی اثر پیشوند و پسوند در شباهت معنایی جملات زبان فارسی با هدف کاربردی در سیستم‌های بازیابی اطلاعات"، سومین کنفرانس داده‌کاوی ایران، تهران: دانشگاه علم و صنعت ایران، 1388.

[10] Williams, N., Zander, S. and Armitage, G., "A preliminary performance comparison of five machine learning algorithms for practical IP traffic flow classification" ACM SIGCOMM Computer Communication Review. Vol. 36, October 2006.

[11] Nguyen, T., T., T. and Armitage, G., "A survey of techniques for internet traffic classification using machine learning" IEEE, Communications Surveys & Tutorials. Vol. 10: 56-76, 2008.

[12] Farid, D., Harbi, N. and Rahman, M., Z., "Combining Naive Bayes and Decision Tree for Adaptive Intrusion Detection" International Journal of Network Security & Its Applications: 12-25, 2010.