New Methods for Text Categorization

HANA KOPACKOVA, LUDEK KOPACEK, RENATA BILKOVA, KAREL NAIMAN
Faculty of Economics and Administration, Institute of System Engineering and Informatics
University of Pardubice
Studentska 84, Pardubice, 53210
CZECH REPUBLIC

Abstract: Text categorization – the assignment of texts to one or more predefined categories based on their content – represents an important component of different information organization and management tasks. Significance of text categorization leads many researchers to find more and more effective methods for this task. It has recently been shown that artificial immune system (AIS) can be successfully used in many machine learning tasks so it can be also used for text categorization task. The aim of this paper is to check applicability of AIS algorithms (Immunos-1, Immunos-2 and Immunos-2A) on text classification task. The results are compared with some classical document classification methods: naïve Bayesian classifier and K-NN classifier.

Key-Words: - text classification, text processing, artificial immune system, Immunos

1 Introduction

Unstructured data (such as text) represent the predominant data type stored online. This implies a necessity of text processing methods use. But the problem with text is that it is not designed for using by computers. Unlike the tabular information typically stored in databases today, documents have only limited internal structure if any. Retrieval of information in unstructured data is then more complicated and it is necessary to find new efficient methods for this task. In this paper we applied those methods: K-nearest neighbor, Naïve Bayes algorithm and Artificial Immune System (AIS) classifier. Forasmuch as classical methods are well known; we will describe only principles of AIS classifier.

2 Document processing methods

There are two types of text processing methods.

Firs can be called linguistic or text understanding and cover natural language processing. People engaged in this kind of research try to design and build software that will analyze, understand, and generate languages that humans use naturally. "Understanding" language means, among other things, knowing what concepts a word or phrase stands for and knowing how to link those concepts together in a meaningful way. This type of research is not the one we chosen for problem solution. While linguistic methods are very difficult so that tools are still not done, statistical methods proved in data mining can find word patterns in large collections of digital documents just now.

Second type can be named as statistical or text classification even if it covers also methods from machine learning. These text mining methods are similar to classical data mining methods, only with transformation of text to standard numerical form. Those methods proved successful without understanding specific properties of text such as the concepts of grammar or the meaning of words. Strictly low-level frequency information is used, such as the number of times a word appears in a document and then methods of machine learning are applied. This approach will be taken as a basis for this paper.

3 Text categorization

The term text categorization covers number of information retrieval tasks that are widely accepted as distinct, but which all involve grouping of textual entities. This entities (documents) can be grouped according to human classifier or some learning algorithm, whose job it is to take training examples and create classifier, which is then able to look at further examples and decide if they fit into the learned concept or not.

Here are some examples of possible usage of text categorization; building of personalized Netnews filter which learns about the news-reading preferences of a user [15], classification of news stories [10] or guidance of a user‘s search on the Web [16], [1], [17], [25]. A growing number of learning algorithms have been applied to text classification task, such as Naive Bayesian [12], Bayesian Network [23], Decision Tree [22], [28], Neural Network [29], Linear Regression [31], k-NN [30], Support Vector Machines [8], [13], Boosting [24] and Genetic Algorithms [21]. A comprehensive comparative evaluation of a wide-range of text classification methods is reported in [8][13].

Applications of artificial immune systems on text categorization were introduced only in small number of publications. The most comprehensive review on the immune system approach to document classification is done in Twycross [26]. Employed algorithm is based on coevolutionary approach described by Potter and de Jong [20] and compared with naïve Bayes algorithm. Test data were obtained from Syskill and Webert Web Page Ratings [18]: 4 datasets contained 2 categories each and about 100 document each.

4 Artificial Immune System classifier

The complexity of many computational problems has led to the development of a range of innovative techniques. One area of research, which has attracted a large amount of interest in recent years, is evolutionary computing. The central idea is the evolution of population of candidate solutions through the application of operators inspired by natural selection and random variation. The humane immune system is an example of a system, which maintains a population of diverse individuals, and has provided the inspiration for a number of artificial evolutionary systems. Immunological metaphors were extracted, simplified and applied to create an effective classification technique.

Mainstream of thoughts about artificial immune systems is concerned on three aspects; immune network, clonal selection, and negative selection. Immune network theory [19], [11] proposes that the immune system maintains a network of cells that learn and maintain memory using feedback mechanism. This theory says that even thou information is learned, it can be forgotten if the information is not intensified. Algorithm iaNet that is based on this theory can be found in [7]. Clonal selection theory [3], [5] is the idea that those cells that are effective at recognizing pathogenic material are selected to survive and propagate. Example of algorithm based on clonal selection theory is CLONALG [6]. Negative selection theory [7], [14] is based on eliminating those cells whose receptors are capable of recognizing self-antigens. This process can be used for anomaly detection algorithms.

The human immune system as a whole consists of a multilayered architecture presenting different types of defence against infectious material called pathogens. The most important layer that inspired birth of artificial immune systems is the third layer – the cellular layer. This layer is composed of variety of different cell types with different roles. Most of the cells belong to the leukocyte family, usually known as white blood cells. These cells (especially B-cells and T-cells) are responsible for anomaly detection, which is proceeded according to affinity [5], [7], [27](degree of similarity between a recognition cell and an antigen). The adaptive ability of the immune system is a process called affinity maturation. During an immune response the recognition cell generate many clones of itself in an attempt to gain a better match next time when the antigen is seen (the process is called clonal expansion). Each clone is then mutated in proportion to the affinity between the recognition cell and the antigen (somatic hypermutation). The last step is called clonal selection and cover elimination of newly differentiated clones carrying low affinity antigenic receptors.

According to Forrest et al. [9] and Twycross [26] some general properties of the immune system can be defined. These properties are diversity, distributed and dynamic nature, error tolerance, self-protection and adaptability. Diversity means that different people are susceptible to different pathogens. Dynamic and distributed nature is given by the number of individual components that are constantly being created and destroyed effecting independently anywhere in the body without central coordination. Error tolerance is based on the fact that the humane immune system makes very few mistakes. The same systems, which protect the body, also protect the immune system itself, due to this fact we can talk about self-protection. Adaptability represents basic presumption for creation of artificial immune systems. The ability to identify and respond to novel pathogens and also retain a memory of past infections started AIS research.

5 Immunos-81

One of the first attempt to use immune system principles as a basis for supervised learning and classification system by Carter [4] was called Immunos-81. The system was designed with the intent of taking advantages of the features of immune systems without remaining too close to the biological aspects. Carter makes the argument that the majority of AIS reviewed at the time adhere too closely to the biological metaphor, which although provides some useful architectural algorithm elements may not necessary be useful from a computational point of view.

The immune system concepts are reduced to their most fundamental level before they are incorporated into the prototype. For example, B cells/antibodies are not randomly generated with a range of binding affinities. Instead, B-cell/antibody generation is under the control solely of entered data.

The algorithm Immunos-81 is a supervised learning system designated for classification from whence it follows that consists of training phase and testing or recognizing phase.

For a consideration of works [4] and [2], Immunos-81 was designed for solving multiple problem domains in parallel that are represented by artificial T-cells. In another words each phase of algorithm (training and recognizing) works in two steps. First step is recognizing the right problem domain. This step is maintained by noticed artificial T-cell and result is the classification of input pattern to problem domain. The second step is particular classification of input pattern into final class in specific problem domain. This two-step architecture can be replaced by several instance of classifier, which each solve one problem domain and algorithm T-cell part can be eliminated [2].

6 Immunos-1

The algorithm Immunos-81 was described in Carter [4], unfortunately the algorithm description was not so detailed and implementation can be speculated. Carter focused his work on description of artificial T-cell, which provides partitioning on right problem domain. How was mentioned this can be let out and each problem domain can be solved by separate instance of algorithm. Although Carters work postulate applicability on multi type antigen vector, description was given only for binary vectors of antigens.

These are the reasons why the result of Immunos-81 is not repeatable. This was confirmed in [2] where the Immunos-81 was reviewed and new version of basic idea of this algorithm was designed. The first two implementations, named Immunos-1 and Immunos-2, were created by Brownlee [2] with the goal of repeating the results of the original work [4]. The third algorithm Immunos-99 was created as the extension of previous version. In next sections are described particular versions of algorithms Immunos-1 and Immunos-2. Detailed information about Immunos-99 can be found in [2].

The Immunos-1 is based on the ideas from Immunos-81 with following differences:

- it solves single problem domain – this mean let out step with T-cells.
- uniform antigen vector structure – each antigen has the same vector length and the same vector structure (order and data type).
- no data reduction – for each antigen in training phase is created one clone of B-cell.
- different way of affinity calculation – in Immunos-81 [4], the affinity values are calculated separately for each paratope (attribute) of antigen and B-cell clone. These affinities are summarized across all B-cells in antigen-group. The avidity between unknown antigen and given antigen-group is the combination of all paratope affinities and concentration of given antigen-group. Avidity calculation in Immunos-1 vide infra.

Training phase consist in division of input antigens into groups per known class label. The B-cell population is created for each class label. No enumeration while training phase is provided.

During classification phase the class label for unknown antigen is looked for. The each B-cell population with known class label competes for unknown antigen. The competition is based on computation of value of avidity between unknown antigen *ag* and all B-cells in each B-cell population.

The term affinity is used in [2] in the wrong meaning. Usually is affinity defined [5], [7], [27] as similarity ratio between B-cell and the antigen so that high value of affinity represent significant similarity of B-cell and antigen conversely low value means weak similarity. In [2] is the affinity used in opposite meaning, which can make confusion. Due to this fact, instead of term affinity we will use the term distance (in the case of nominal attributes will be used the term general distance).

Unknown antigen is classified into the appropriate class of B-cell population with the highest value of avidity (1).

$$class = \arg\max_i (avidity(ag, i)) \quad (1)$$

$$avidity(ag,i) = \frac{N_i}{\sum_{j=1}^{N_i} D(ag, ab_j)} \quad (2)$$

Avidity (*avidity*(*ag*, *i*)) between unknown *ag* and *i*-th B-cell population are computed by equation (2) where N_i is number of B-cells in the *i*-th population and $D(ag, ab_j)$ is distance between unknown antigen *ag* and *j*-th B-cell from *i*-th B-cell population.

$$D(ag, ab_j) = \sqrt{\sum_{i=1}^{Na} d_i(ag, ab_j)} \quad (3)$$

- where $d_i(ag,ab_j)$ for numeric attributes are computed by equation (4) and for single nominal attributes by equation (5). N_a is number of attributes of antigen *ag*.

$$d_i(ag, ab_j) = (ag_i - ab_{i,j})^2 \quad (4)$$

$$d_i(ag, ab_j) = \begin{cases} 0 & if \quad ag_i \equiv ab_{i,j} \\ 1 & if \quad ag_i \neq ab_{i,j} \end{cases} \quad (5)$$

- where ag_i is i-th attribute of antigen *ag* and $ab_{i,j}$ is *i*-th attribute of *j*-th B-cell. Lets make a remark; $D(ag, ab_j)$ according to (2) in the [2] is improperly called affinity $affinity(ag, ab_j)$.

Avidity, in equation (2), means inverse value of average distance between vector of unknown antigen and all B-cell vectors in given B-cell population.

7 Immunos-2 and Immunos-2A

Distinction between Immunos-1 and Immunos-2 is that the Immunos-2 provides full data reduction in training phase. In Immunos-2 are whole B-cell populations reduced after creation of one B-cell taking mean attribute value across the entire population. In case of nominal attributes, the value with highest frequency is taken. Each B-cell population is represented by one exemplar.

Classification is proceeded in similar manner as in Immunos-1. Affinity is calculated between unknown antigen and prepared exemplar for each group. Unknown antigen is classified into the appropriate class of B-cell population with the highest value of avidity.

Avidity calculation in Immunos-2 [2] is done according to the same equation (2) as in Immunos-1, it means that affinity calculated between unknown antigen and prepared exemplar for each group is furthermore weighted by the number of B-cells in particular class. This type of classification prefers highly populated classes of B-cells in contrast to classes with only few B-cells. In case of over different classes may happen this possible situation: big group of B-cells is selected as similar although this class has smaller similarity ratio in comparison with the small one. We can sum this approach as being dependent on apriory probability of pattern occurrence in particular group. Nevertheless the question is what will be the real impact on classification. For the comparison we designed new version of algorithm Immunos-2 without weighting affinity by the number of B-cells in particular class. This version was named Immunos-2A (for lack of better name).

8 Experimental results

Experiments are conducted on the 20 Newsgroup data set. Several subsets of documents with various numbers of documents are chosen. The details of the subsets are given in Table 1.

Table 1: Details on testing subsets

Datasets	Description	Total # of docs
Subset_1	Selection of first 100 documents from each class	2000
Subset_2	Random selection of 100 Documents from each class	2000
Subset_3	Random selection of 150 documents from each class	3000
Subset_4	6 categories per 50 documents and 14 categories per 100 documents	1700
Subset_5	10 categories per 50 documents and 10 categories per 100 documents	1500

Test 1: cover comparison of Immunos-1, Immunos-2, Immunos-2A, naive Bayes and K-NN on the subset_1, subset_2 and subset_3.

Test 2: cover comparison of Immunos-2, Immunos-2A, naive Bayes and K-NN on all subsets.

All versions of algorithm Immunos, in the TEST-1, achieve better results then naïve Bayes or K-NN.

Results of TEST-2 show evident impact of weighting by the number of B-cells in particular class. For subsets, which have in all classes the same number of patterns, both algorithms Immunos-2 and Immunos-2A achieve similar results. Classes from subsets with different number of patterns are more problematic and Immunos-2 did not succeeded as well as Immunos-2A. The result can be seen in figure 1.

Table 2: Experimental results.

	Immunos1 (IM-1)	Immunos-2 (IM-2)	Immunos-2A (IM-2A)	Naive Bayes (NB)	1-NN	5-NN
Subset_1	67,5%	75,9%	76,5%	54,8%	39,1%	35%
Subset_2	54,6%	73,5%	73,5%	49,8%	26,6%	23,7%
Subset_3	56,7%	71,8%	71,9%	51,6%	32,3%	29,2%
Subset_4	55,7%	64,4%	77,7%	54,9%	40%	34,6%
Subset_5	48,9%	53,3%	76,7%	56,5%	40,5%	36,5%

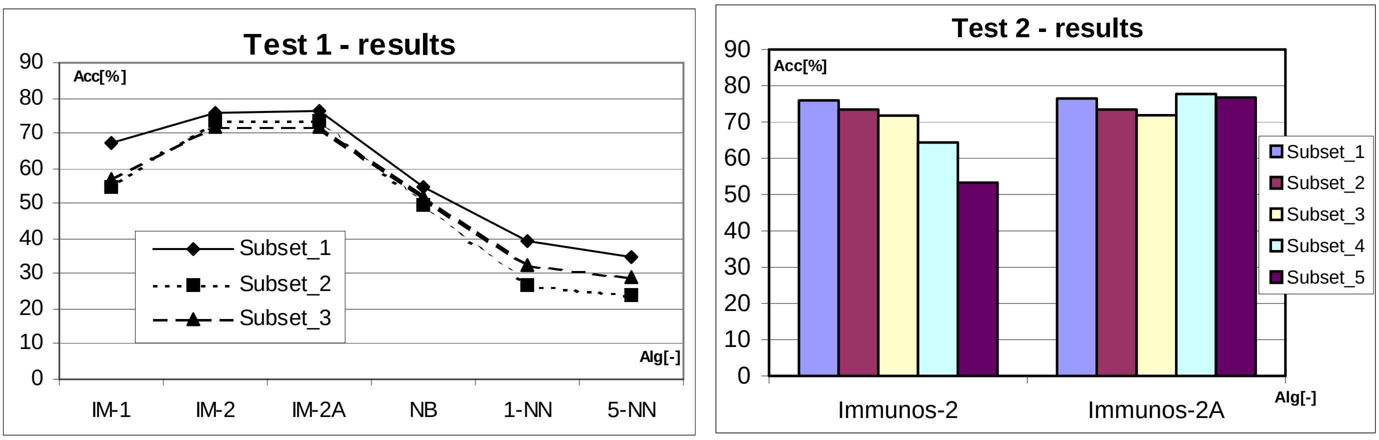

Figure 1: Results of Test-1 and Test-2.

9 Conclusion

In this paper was discussed applicability of artificial immune systems on text classification task. Explanation was given for algorithms Immunos-1, Immunos-2 (both based on Immunos-81) and Immunos-2A, which represent adaptation of algorithm Immunos-2. Testing was done on all mentioned algorithms and compared with classical methods like naïve Bayes a K-NN. Testing datasets were prepared from dataset 20NewsGroup. The results show applicability off all introduced algorithms nevertheless algorithm Immunos-2A (prepared especially for these experiments) gave the best and the most stable results.

10 Acknowledgement

This research and paper was created with a kind support of the Grant Agency of the Czech Republic, grant number GACR 402/05/P155.

References:

[1] Bollacker K., Lawrec S., Giles L., Citeseer: An Autonomous System for Processing and Organizing Scientific Literature on the Web, *Conference on Automated Learning and Discovery*, Pittsburgh, 1998

[2] Browlee, J. *Immunos-81, The Misunderstood Artificial Immune System*. Technical Report, no. 3-01, January 2005.

[3] Burnet, F. M. *The Clonal Selection Theory of Immunity*. Vanderbilt University Press, Nashville, TN, 1959.

[4] Carter, J. H., The Immune system as a model for classification and pattern recognition. *Journal of the American Informatics Association*, vol. 7, 2000.

[5] de Castro, L. N. - Von Zuben, F. J., "The Clonal Selection Algorithm with Engineering Applications", In Proceedings of GECCO'00, Workshop on Artificial Immune Systems and Their Applications, pp. 36-37, 2000.

[6] de Castro, L. N. - Zuben, F. J. Learning and Optimization Using the Clonal Selection Principle. *IEEE Transactions on Evolutionary Computation, Special Issue on Artificial Immune Systems*. 2002, vol. 6, no. 3, pp. 239-251.

[7] de Castro, L. N. - Timmis, J. I. *Artificial Immune Systems: A New Computational Intelligence Approach*, Springer-Verlag, London, September, 357 p., 2002.

[8] Dumais S. et. al., Inductive learning algorithms and representations for text categorization. Proceedings of the *7th International Conference on Information and Knowledge Management (CIKM 98)*, 1998.

[9] Forrest, S., Hofmeyr, S. A. Immunology as information processing. In *Design Principles for Immune Systems and Other Distributed Autonomous Systems*. Oxford University Press, New York, 2001.

[10] Hayes P. et al., A news story categorization system, *Second Conference on Applied Natural Language Processing*, 1988

[11] Jerne, N. K. Towards a Network theory of the Immune System. *Annals of Immunology*, 125c:373–389, 1974.

[12] Joachims T., A probabilistic analysis of the Rocchio algorithm with TFIDF for text categorization, Proceedings of ICML-97, *14th International Conference on Machine Learning*, 1997.

[13] Joachims T., Text categorization with support vector machines: learning with many relevant features. Proceedings of *ECML-98, 10th European Conference on Machine Learning*, 1998.

[14] Kim, J. - Bentley, P. “The Artificial Immune Model for Network Intrusion Detection, *7th*

European Conference on Intelligent Techniques and Soft Computing EUFIT'99), Aachen, Germany, 1999.

[15] Lang K., NewsWeeder: Learning to Filter Netnews, *International Conference on Machine Learning*, 1995.

[16] Mitchell T. et al., WebWatcher: A Learning Apprentice for the World Wide Web, *AAAI Sprig Symposium on Information Gathering from Heterogenous, Distributed Environments*, 1995

[17] Mladenic D.: Personal WebWatcher: Implementation and Design, *Tech. Report IJS-DP-7472*, J. Stefan Inst., 1996

[18] Pazzani, M. Syskill and Webert web page ratings. UCI Repository of Machine Learning Databases, Department of Information and Computer Sciences, University of California, Irvine, CA. <http://www.ics.uci.edu/~mlearn/MLRepository.html>

[19] Perelson, A. S. Immune Network Theory. Immunol. Rev., 110 (1989) 5-36.

[20] Potter, M. A., De Jong, K. A. The coevoution of antibodies for concept learning. In Proceeding of the Fifth International Conference on Parallel Problem Solving from Nature, Springer-Verlag, Amsterodam, 1998, p. 530-539.

[21] Prasanna K., Khemani D. Applying Set Covering Problem in Instance Set Reduction for Machine Learning Algorithms. WSEAS Multiconference: Software Engineering, Parallel & Distributed Systems (SEPADS 2004). Salzburg, Austria, 2004.

[22] Quinlan J. R., C4*.5: Programs for machine learning*, Morgan Kaufmann, 1993.

[23] Sahami M., Learning limited dependence Bayesian classifier. In KDD- 96: Proceedings of the second international *Conference on Knowledge Discovery and Data Mining*, AAAI press, 335-338, 1996.

[24] Schapire R. E., Singer Y., Boostexter: A boosting-based system for text categorization. *Machine Learning*, 39(2/3), 2000.

[25] Sklenák V. a kol., *Data, informace, znalosti a internet*, 1. vyd. Praha, C. H. Beck, 2001

[26] Twycross, J. *An Immune System Approach to Document Classification.* Master Thesis, University of Sussex, Hewlet-Pacard Research Labs, Bristol, 2002.

[27] Watkins, A. – Boggess, L. A New Classifier Based on Resource Limited Artificial Immune Systems. In Proceedings of *Congress on Evolutionary Computation, Part of the 2002 IEEE World Congress on Computational Intelligence held in Honolulu*. May 2002, HI, USA, pp. 1546-1551.

[28] Weiss S. M et. al., *Maximizing text-mining performance*, IEEE Intelligent systems,1999.

[29] Wiener E., Pederson J. O., Weigend A. S., A neural network approach to topic spotting. Proceedings of *SDAIR-95, 4th Annual Symposium on Document Analysis and Information Retrieval*, 1995.

[30] Yang Y., An evaluation of statistical approaches to text categorization. *Journal of Information Retrieval*, Vol.1, No.1/2, 1999.

[31] Yang Y., Chute C. G., A linear least squares fit mapping method for information retrieval from natural language texts. Proceedings of the *14th International Conference on Computational Linguistics (COLING 92)*, 1992.