Text-based Decision Making with Artificial Immune Systems

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

Abstract: Decision making is one of the most important manager activities. Especially in these days, full of different information, it is necessary to distinguish between important and unimportant information to make significant decision. Especially unstructured documents can be full of hidden information. Usage of text classification can significantly lower manager workload and raise objectivity of decision making process. Automated processing of text documents can also prevent simplification and generalisation, which allow us to decide on the base of small amount of cases and widen this decision on all cases. The aim of this paper is to find methods which fulfil criteria put on text classification done by managers due decision making process.

Key-Words: - Decision making, Management, Text classification, Artificial immune system, Immunos-81

1 Introduction

Proces of decision making represents complex and difficult problem, which is highely dependant on the accessible information. Seeing that the amount of produced information is growing faster than the ability of information consumers to find, retrieve and use this information, decision makers must solve this situation. It is quite easy to find many methods that makes access to numerical data. For example database software, spreadsheets or statistic packages. But it is much harder to provide effective access to textual data that consists of expresions in human language, because of the relationship between its form and its content, which is less clear than in numeric data.

2 Proces of decision making

We can find three main types of decision making: intuitive, rational and limited rationality. Intuitive decision making is used when fast solution is needed. The act of decision-making is done by individual, mostly using intuition and previous experience. As an advantages of intuitive decision making we can find speed and interest of decision maker. Nevertheless emotion based intuitive decision-making can also have some serious disadvantages. Failure to consider other available options, inaccurate or irrelevant information used for decision and finally, intuitive reasoning is problematic in group situations where decisions need to be made collectively.

Rational decision making model assumes that decision maker posses a utility function (an ordering by preference among all the possible outcomes of choice), that all the alternatives among which choice could be made were known, and that the consequences of choosing each alternative could be ascertained [1], [2]. Although the assumptions of rational model (results and alternatives are ascertained, information are complete and so on) cannot be satisfied even remotely for most complex situations in the real world, they may be satisfied approximately in some isolated problem situations without uncertainty.

Although examples of rational decision-making process are not rare, limits of incomplete information, inconsistency, and institutional constraints on alternatives brings us to the models of limited (bounded) rationality. These models retain the same process of decision-making but incorporate many additional limitations. Simon [3] describes ‘satisfying’ behavior - a decision making process which searches for ‘good enough’ options, rather than an optimum solution. With satisfying decision making becomes something which is carried out in a limited time, with limited information and with some limits on the individuals concerned.

3 Textual information in decision making

Problem of textual information is considered by Mark Tucker in [4]. ”The ratio of unstructured to structured information in most organizations is easily 9 to 1, yet many of us spent most of our time worrying about – indeed, dedicating our careers to – managing the most familiar 10 percent of the problem: structured information… Business processes have always relied on unstructured information, and the volume and sources of this information are increasing, not decreasing. The World Wide Web, the corporate Intranet, email, and online discussion groups are just a few of the familiar examples… Information drives our business decisions,

and it always has. What has changed dramatically is the kind of decisions we make. As the economy shifts from an industrial model to a knowledge-driven one, we need more and more information to support the decision-making process, and the dynamic nature of our business environment is such that less and less of this information fits the structured information model."

Forest Research [5] has predicted that unstructured data (such as text) will become the predominant data type stored online. This implies a great opportunity of possible more effective use of repositories of business communications, and other unstructured data, by using computer analysis. 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. Furthermore, the important information they contain is not explicit but is implicit: buried in the text. [6]

Managerial decision-making process can be highly dependent upon hidden information in text documents nevertheless careful reading and sorting of documents is time consuming work. This type of activity waste working time of managers and at the end, it can cause wrong decision.

Now we must ask how we can help managers to cope with numerous text documents.

The aim of this paper is finding of methods which can fasten decision-making process on all levels of management and make it much easier for managers using great amount of information in text form. Usage of text categorization methods, which are well known in the branch of information retrieval, but they are not often used to support decision making process can significantly lower manager workload. Another aim can be defined as raise of objectivity in decision-making process. Automated processing of text documents can prevent simplifications and generalisation, which allow us to decide on the base of small amount of cases and widen this decision on all cases. Unfortunately, this approach is commonly used having too much text documents and only little time to read them.

4 Text 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.

5 Text classification

The term text classification cover 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 classification; building of personalised Netnews filter which learns about the news-reading preferences of a user [7], classification of news stories [8] or guidance of a user's search on the Web [9, 10, 11, 12].

A growing number of learning algorithms have been applied to text classification task, such as Naive Bayesian [13], Bayesian Network [14], Decision Tree [15] [16], Neural Network [17], Linear Regression [18], k-NN [19], Support Vector Machines [20, 21], Boosting [22] and Genetic Algorithms [23]. A comprehensive comparative evaluation of a wide-range of text classification methods is reported in [20, 21].

5.1 Assumptions for application of text classification in decision making process

Learning algorithms for text classification were usually tested in specific environment which cover:

- great amount of training documents (Reuters collection, newsgroups, TDT Pilot study corpus... – about 20000 documents),
- and approximately similar length of documents.

These conditions are almost impossible to guarantee in real managerial decision-making. The aim of this paper and the whole research is to find such methods, which can be applied directly by managers if they need them. This statement assumes that managers are the persons who will decide which documents will be used as

training documents. No one can expect that he/she will have great number of training documents with similar length. Probably they will use maximum tens of documents.

Methods suitable to support managerial decision-making process must fulfil these criteria:

- easy to implement,
- fast to process categorization,
- cheap,
- stable for differences in length of document,
- learning model build from small number of documents,
- high precision.

Two classical methods and one very new method will be tested for being used in managerial decision-making. These methods are K-nearest neighbour, Naive Bayes algorithm and Artificial Immune System (AIS) classifier. Forasmuch as classical methods are well known, we will describe only principles of AIS classifier.

6 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.

Mainstream of thoughts about artificial immune systems concerns on three aspects; immune network, clonal selection, and negative selection. Immune network theory [24, 25] 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. Clonal selection theory [26, 27] is the idea that those cells that are effective at recognising pathogenic material are selected to survive and propagate. Negative selection theory [28, 29] 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 (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. [30] and Twycross [31] 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 immnue system makes very few mistakes. The same system 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.

6.1 Immunos-81

The first attempt to use immune system principles as a basis for supervised learning and classification system by Carter [32] 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 made 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 primary reasons why the immune system was selected as the basis for a classification algorithm was

given its ability to readily accept patterns of arbitrary length, and its ability to maintain and exploit previously learned information efficiently for improved performance in future encounters. It was shown to provide good results in terms of classification accuracy on the Cleveland heart disease datasets.

The terminology used by Carter and explained by Brownlee [33] is described in next section.

T-cell – These elements learn the primary structure of antigens in the system and decide what is interesting to the system. The T-cells act as gatekeepers to the system, both partitioning learned information and deciding how new information is exposed to the system. One T-cell exists for each antigen-type in the system, and each T-cell has one or more clone (groups) of B-cells.

B-cell – These elements learn the surface features of antigens. In the system, B-cells represent an instance of an antigen-type, more specifically an instance of an antigengroup. It is mentioned by Carter that B-cells are not represented explicitly by Immunos- 81, rather they are incorporated into a clone or group structure.

Antigen – An antigen is a molecule that elicits an immune response, which in this case is represented by a single training data instance. An antigen is represented as a data vector of attribute or field values where the nature of each attribute (name, data type) are known. Antigens can be variable length, and can have missing field values. Further, Immunos-81 supports the acceptance of antigens from differing problem domains.

Antigen-type – Confusingly called antigen-class by Carter, the antigen-type refers to a specific problem domain identified by a specific name and series of attributes. This is the primary structure learned by T-cells.

Antigen-group (clone) – An antigen group is a set of antigens of a specific antigen-type that are of interest. Given that the nature of the problem addressed by Immunos-81 is that of classification, an antigen-group represents all antigens with a specific classification label. Represented in the system, this group is called a clone of B-cells controlled by a single T-cell.

Concentration (antigenicity) – An interesting term used to describe the number of antigens (training instances) with a specific class label (an antigen-group) was antigen concentration or antigenicity. This concentration obviously indicates the amount of attention each antigen-group will receive, as well as aids in defining the size of the antigen-groups clone of B-cells.

Paratope and Epitope – A paratope is a part of the receptor on an antibody molecule that binds to a matching part of the antigen molecule called the epitope. In Immunos-81 these terms are used to described individual attributes on the recognition cells (T-cells and B-cells), and antigens respectively.

Affinity – A recognition cell has specificity with an antigen, the degree that is called affinity. In the Immunos-81 system, affinity is a match between a similarity/dissimilarity measure between a T-cell and an antigen, and a B-cell and an antigen.

Avidity – Avidity is the term used to describe the sum affinities for all B-cells of a clone for a given antigen. This value is used to allow clones of a T-cell to compete with each other to classify an antigen.

Amino Acid Library (AALib) – As mentioned, antigens that exhibit the largest immune response consists of proteins, which are sequences of amino acids. Immunos-81 uses a similar concept to manage antigen details. A given-antigen type has a name (problem domain name) and a series of attributes (order, attribute names, data types, etc.). This information is required for a T-cell and throughout the learning process. All this domain knowledge is centralised and stored in an AALib construct.

7 Experimental results

In the testing environment we used 50 documents in Czech language; 25 of them were focused directly on the branch of waste management and the rest 25 documents were not specialised – only covered problem of environment. The shortest document had only 98 words and the longest had 1400 words. For feature selection were used five different methods: Term frequency selection (selected were features that occurred in the document more then once), Document frequency (selected were features that occurred in more then one document), Chi-square, Mutual information, and Information gain [34]. Two methods for term weighting were tested TF – term frequency and TFIDF. After pre-processing stage database was filled with 10571 words.

The result can be seen in figure 1 (CCI means correctly classified instances and K means Kappa statistics). Experiments that were published for example in [19] introduced K-NN method as very efficient, however these experiments used Reuters corpus filled with articles of similar legth. Our experiments proved that K nearest neighbour algorithm is very sensitive to length differences (documents using same words have long Euclidean distance between them if they differ in length) so it is not suitable to support managerial decision-making as it was described here. On the other hand Naive Bayes and Artificial Immune Systems can serve as very helpful tool.

Differences between TF and TFIDF methods are not so dominant to say that one of them is better but it can be said that TF method is cheaper and sufficing.

	Naive Bayes		K-NN		Imunos-81	
	CCI-NB [%]	K-NB [%]	CCI-KNN [%]	K-KNN [%]	CCI-81 [%]	K-81 [%]
TF-TF	88,00	76,00	56,00	12,00	100,00	100,00
TF-DF	90,00	80,00	56,00	12,00	100,00	100,00
TFchi-square	96,00	92,00	62,00	24,00	100,00	100,00
TFinformation gain	98,00	96,00	62,00	24,00	100,00	100,00
TFmutual information	96,00	92,00	50,00	0,00	100,00	100,00
TFIDF – TF	90,00	80,00	56,00	12,00	96,00	92,00
TFIDF-DF	92,00	84,00	58,00	16,00	98,00	96,00
TFIDFchi-square	96,00	92,00	66,00	32,00	100,00	100,00
TFIDFinformation gain	96,00	92,00	68,00	36,00	100,00	100,00
TFIDFmutual information	92,00	84,00	50,00	0,00	96,00	92,00

Figure 1: Experimental results

8 Conclusion

Unstructured information in the form of textual documents are used in managerial-decision making very often, nevertheless support for this kind of action is inadequate. In this paper we shortly introduce tools that deal with textual information including proposal of methods that fulfil demands specially put on this kind of decision-making.

9 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] H. A Simon, A Behavioral Model of Rational Choice. Cowle: Foundation Paper 98. *The Quarterly Journal of Economics*, vol. LXIX, February 1955.

[2] H. A Simon, *Report of the Research Briefing Panel on Decision Making and Problem Solving.* National Academy of Sciences. Washington, DC: National Academy Press, 1986

[3] H. A. Simon, *Models of Bounded Rationality*., Cambridge, M.A.: Harper and Row, 1983.

[4] Tucker, M. Dark Matter of Decision Making. Intelligent Enterprise Magazine, vol. 2, num. 13. 1999

[5] Forrester Research. Coping with Complex Data. The Forrester Report, 1995

[6] Tkach, D. Text Mining Technology, Turning Information Into Knowledge. White paper from IBM. 1998

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

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

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

[10]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

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

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

[13] 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.

[14] 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.

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

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

[17] 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.

[18] 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.

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

[20] 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.

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

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

[23] 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.

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

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

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

[27] 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.

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

[29] 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.

[30] 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.

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

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

[33] Brownlee, J. The misunderstood artificial immune system. Technical report No. 3-01. January, 2005.

[34] Janakova, H. Text categorization with feature dictionary – problem of Czech language. WSEAS TRANSACTIONS on INFORMATION SCIENCE AND APPLICATIONS, Issue 1, Volume 1, July 2004, s. 368 - 372, ISSN 1790-0832