ISSN:0975-9646
Swati Wakode et al, / (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 6 (3) , 2015, 2868-2870

Classification of Unstructured Data using Soft Computing: A Survey

Swati Wakode [1], Rakesh Mallesh [1], Mandar Wagh [1], Manisha P Mali [2]

[1]*U.G.Student* [2]*Professor*
[1,2]*Department of Computer Engineering*
[1,2,]*Vishwakarma Institute of Information Technology, Pune*

Abstract - Vast amounts of new information and data are generated everyday through economic, academic and social activities in digital media, much with significant potential economic and social value. In some applications like spam filtering and language identification, semantic analysis there is a need for this data is to be classified depending on its categories. This article provides a survey of the available literature on text classification. The different soft computing approaches are discussed. The utility of these soft computing approaches is highlighted.

Index Terms -Fuzzy logic, Genetic algorithm, Neural Network, Rough sets, classification.

1. INTRODUCTION

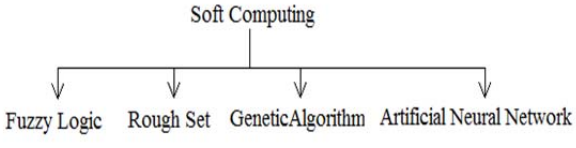
Unstructured data refers to information that is not organized in a pre-defined manner. From [1] we studied that unstructured data cannot be stored in rows and columns in relational database. An example of unstructured data is a document that is archived in a file folder. The advantage of unstructured data is that no additional information is necessary for classification. The problem in library science, information science and computer science is document classification or categorization. Assigning a document to one or more classes or categories is the task. This may be done "manually" or "algorithmically". If we think of categorizing data manually it is a hectic job and time consuming .Even it requires much cost. For example, Mayo Clinic it requires $1.4 million annually for coding patient-record events. This is a high amount. Data Mining helps in classifying documents automatically into predefined classes based on their content. Many algorithms have been developed to deal with automatic text classification.

Some of the key methods are given in [3], which are commonly used for text classification, are Decision Trees, Pattern (Rule)-based Classifier, SVM Classifiers, Bayesian (Generative) Classifiers, and Neural Network Classifiers. Classification Using Decision Tree can be done in sequential or parallel way. A decision tree has internal node and leaves. Internal node has decision associated with it and leaves has class label attached to it. Bayesian classification is based on Bayes theorem. Naïve Bayes classifier assumes that the effect of an attribute value on a given class is independent of the values of the other attributes. It is made to simplify the computations involved and, in this sense, is considered “naïve”. Neural Network is another approach used for classification. According to [3] Neural Network performs well with missing or incomplete data and also it does not make any assumptions about nature of distribution of the data. Artificial Neural network is one of the different approaches of soft computing. Some other soft computing approaches are Fuzzy Logic (FL), Genetic Algorithm (GA), and Rough Set (RS).

The steps given in [4] for Text Classification process involves Document Collection, Preprocessing, Feature Selection and Finally Classification. Data in the real word is dirty. Incomplete data, noisy data, inconsistent data is referred to as dirty data. Quality decisions are based on quality data. So there is need to perform text preprocessing as duplicate or missing data may cause incorrect or even misleading statistics. The main objective of preprocessing is to obtain the key features or key terms from text documents and also to enhance the relevancy between word and document and the relevancy between word and category. In [5] the main preprocessing steps are described and these steps are stop word removal, upper case to lower case, stemming, and encoding. Next step is Feature Selection. It involves selecting subset of features from the original documents and constructing vector space model [4]. Using Feature Set, Classification is done.

2. SOFT COMPUTING

Soft computing is a term applied to a field within computer science. It is characterized by the use of inexact solutions to computationally hard tasks such as the solution of NP-complete problems. There is no known algorithm that can compute an exact solution in polynomial time (i.e. the solution may be 0, 1 or in between 0 and 1). Soft computing is tolerant of imprecision, uncertainty, partial truth, and approximation, unlike hard computing. In effect, the human mind is the role model for soft computing.



Soft Computing approaches are Fuzzy Logic, Rough Set, Genetic Algorithm and Artificial Neural Networks,

A. Fuzzy Logic

Fuzzy Logic was introduced in year 1965 by Lotfi A. Zadeh, Ph.D., University of California, Berkley. Fuzzy logic is an approach to computing based on 'degree of truth' rather than 'True or False'. The fuzzy systems convert these rules to their mathematical equivalents. The job of the system designer and the computer is simplified which results in accurate representations of the way systems behave in the real world. The value 0 and 1 describes ‘not belonging to’ and ‘belonging to’ a conventional set, respectively. While the values between 0 and 1 as described in [7] are represented as ‘fuzziness’.

Fuzzy sets [6] provide a robust mathematical framework for dealing with “real-world” imprecision and non-statistical uncertainty. An important property of fuzzy set is that it allows partial membership. A fuzzy set is a set having the degrees of membership between 0 and 1. The operations carried out on fuzzy set are all set operations such as Union, Intersection, Complement, etc. The set operations are shown by Venn diagram.

Advantages

- Simplicity and flexibility.
- It also allows vague linguistic terms in the rules.

Disadvantage

- It is difficult to identify membership function.

B. Rough Set

In Rough Set, membership is not the primary concept as that of Fuzzy Logic. Rough Set represents a different mathematical approach to vagueness and uncertainty. Rough Set is used to discover the data dependency, to evaluate the importance of attribute, to recognize the patterns.

As described in [8] the objects are represented in table called information table in which the rows represents object and column represent object features. For example, if objects are patients suffering from a certain disease than symptoms of the disease form information about patients.

Advantages

- Provides efficient algorithms for finding hidden patterns in data.
- Quantitative and Qualitative data is allowed by it.
- Finds minimal sets of data (data reduction).
- Evaluates significance of data.
- Generates sets of decision rules from data.
- Offers straightforward interpretation of obtained results.

C. Genetic Algorithm

Genetic algorithms belong to the larger class of Evolutionary Algorithms (EA), which generate solutions to optimize problems using techniques inspired by natural evolution. Genetic algorithm requires less information about the problem and uses rule based classifiers in general. It aims to use selective `breeding' of the solutions to produce `offspring' better than the parents by combining information from the chromosomes.

The evolution usually starts from a population (features) generated by preprocessing the documents and is an iterative process, each iteration is called a generation. In each generation, the fitness score of every individual in the population is evaluated. A fitness score is used to represent an individual's ability to survive in the next generation. An individual is any possible solution. The more fit individuals are selected from the current population, and each individual is modified to form a new generation. The new generation produced is used for next iteration of algorithm. The iteration process will terminate when the desired number of individuals are selected.

Each individual is represented in an array of 0's and 1's, which is called as chromosomes, where each 0 and 1 represent characteristics of the individual. After initial population is generated the algorithm performs following operations on the population.

- Fitness: It is used to evaluate Fitness score of an individual (feature). With the help of this score the selection process will be carried out
- Selection: It is used to select the chromosome from the population by using the fitness score.
- Crossover: Two individuals are chosen from the population using the selection operation. A position is selected in the chromosome. The bits lying after the selected position in the parent are swapped so as to produce new offspring.
- Mutation: Each new offspring is used and one bit or more than one bit will be flipped (1 becomes 0 and 0 becomes 1). It is used to maintain diversity within the population.

Advantages

- It can solve with multiple solutions.
- Every optimization problem which can be described with the chromosome encoding could be solved by it.
- It is easy to understand and it practically does not demand the knowledge of mathematics.

Disadvantages

- The population has lot of subjects which makes it difficult to the genetic algorithm to find a global optimum.
- It cannot assure constant optimization response time.

D. Artificial Neural Network

Artificial neural networks (ANNs) are computational models inspired the brain, and are used to estimate or approximate functions that can depend on a large number of inputs and are generally unknown. Artificial neural networks are generally presented as systems of interconnected "neurons" which can compute values from inputs [10].

Advantages

- They can use it for non-linear problems.
- Using Back propagation learning algorithm is widely used in solving various classifications and forecasting problems.

Disadvantages

- It behaves as Black box system due to which the user cannot explain how learning from input data was performed.

3. CONCLUSION

Soft computing methods are effective in search and optimization when the underlying search space is large, multimodal, and when the characteristics of the search space are not well understood. The memory and computational time requirements for soft computing approaches are higher than for traditional techniques. Fuzzy Sets are suitable for handling the issues related to understandability of patterns, incomplete or noisy data and human interaction and can provide approximate solutions faster. Neural networks are nonparametric, robust. Good learning and generalization capabilities in data-rich environments are exhibited by neural networks. Efficient search algorithms to select a model based on some preference criterion/objective function from mixed media data is provided by it. Different types of uncertainty in data is handled by rough sets. Along with these approaches we can also use hybrid approaches. Hybridization of these approaches is important because of their capabilities in handling several real world problems involving complexity, noisy environment, imprecision, uncertainty, and vagueness. For example, advantages of massive parallelism, robustness and learning in data rich environment is provided by Neuro-Fuzzy computing, GA-ANN can be used as a hybrid approach for text classification.

REFERENCES

[1] olf Sint , Sebastian Schaffert , Stephanie Stroka and Roland Ferstl,"Combining Unstructured, Fully Structured and Semi-Structured Information in Semantic Wikis"

[2] Canasai Kruengkrai, Chuleerat Jaruskulchai, “A Parallel Learning Algorithm for Text Classification,” The Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-2002), Canada, July 2002.

[3] Mrs. Sayantani Ghosh , Mr. Sudipta Roy, and Prof. Samir K. Bandyopadhyay, “A tutorial review on Text Mining Algorithms”, International Journal of Advanced Research in Computer and Communication Engineering Vol. 1, Issue 4, June 2012.

[4] Bhumika, Prof Sukhjit Singh Sehra, Prof Anand Nayyar,“A Review Paper on Algorithms used for Text Classification”, International Journal of Application or Innovation in Engineering & Management (IJAIEM) Volume 2, Issue 3, March 2013.

[5] Francesca Possemato and Antonello Rizzi, Member, IEEE, “Automatic Text Categorization by a Granular Computing Approach: facing Unbalanced Data Sets”.

[6] Sumit Ghosh, Qutaiba Razouqi,H. Jerry Schumacher, and Aivars Celmins “A Survey of Recent Advances in Fuzzy Logic in Telecommunications Networks and New Challenges”IEEE transactions on fuzzy systems, vol. 6, no. 3, August 1998 (443-447)

[7] Sivanandam S. N. (2008) “Principles of Soft Computing”, New Delhi, Ar Emm International

[8] Aboul Ella HASSANIEN, Jafar M.H. ALI “Rough Set Approach for Generation ofClassification Rules of Breast Cancer Data”,Institute of Mathematics and Informatics, Vilnius 2004 Vol. 15, No. 1, (23–38)

[9] Shivani Patel, Prof. Purnima Gandhi,“A detailed study on Text Mining using Genetic Algorithm”, ISSN: 2321-9939 International Journal of Engineering Development and Research IJEDR (101-105)

[10] http://en.wikipedia.org/wiki/Artificial_neural_network

[11] Sushmita Mitra, Sankar K. Pal, Pabitra Mitra, “Data mining in Soft Computing Framework: A Survey”, IEEE transactions on neural networks, vol. 13, no. 1, January 2002