Grzegorz BOROWIK
WUT, INSTITUTE OF TELECOMMUNICATIONS, Nowowiejska 15/19, 00-665, Warsaw

Application of Boolean Function Complementation in Data Mining Algorithms

Grzegorz BOROWIK, PhD, Eng.
earned his Master's in mathematics and PhD in telecommunications from Warsaw University of Technology in 2002 and 2007 respectively. He joined the Institute of Telecommunications of WUT in 2003, where he is an Assistant Professor and Managing Editor of International Journal of Electronics and Telecommunications. His research interests include Knowledge Discovery and Data Mining, Optimization Techniques, Logic Synthesis, Design, and Testability of PLDs, Cryptography. Dr. Borowik has authored or coauthored about 50 scientific papers published in referred Polish and international scientific journals as well as conference proceedings.

e-mail: G.Borowik@tele.pw.edu.pl

Abstract

The paper presents a method which supports three main tasks of data mining algorithms such as feature extraction, rule induction, and data discretization. It have been proved that the problems can be reduced to very efficient unate complementation algorithm. That algorithm is based on recursive execution of Shannon expansion procedure. It continues until at each leaf of the recursion tree yields the data which can be easily complemented. The final result is obtained merging the results in the subtrees. According to the results of computer-based experiments this algorithm has proved very efficient.

Keywords: feature extraction, rule induction, discretization, quantization, data mining, Boolean function complementation, multimodal data, telecommunications, biomedical engineering.

1. Introduction

Nowadays huge amounts of useful data are being collected. The dramatic increase of data volume causes difficulties to extract useful information for decision support, discovery of underlying principles, or different analysis tasks. It yields a large demand for methods that deal with the increasing number of database records and attributes.

The core of this research is a novel approach to knowledge discovery, as well as application of unique methodology based on unate complementation algorithm. The problem is located at the intersection of the fields of logic synthesis (logic circuit design and optimization) and knowledge discovery, where Shannon expansion and other specific methods used in logic design are applied to resolve the task of feature extraction, rule induction, and data discretization.

One of the major application area of the proposed solution is telecommunications, especially anomaly detection in telecommunications networks and systems. Since the decision on anomaly detection is made using a collection of decision rules induced by the algorithm for the training data, the algorithm is the standard procedure of machine learning. The system creates a knowledge base containing patterns of analyzed anomalies. Then, using the algorithm of decision-making and classification, it creates a set of decision rules classifying the current data. It should be noticed that emails can contain text, pictures and other multimedia (multimodal data). A characteristic example of training data is the database for e-mail classification [11], which contains 58,042 records represented by 64 attributes, for which the objective of the algorithm is to obtain decision rules classifying data according to the following conditions: y_spam, n_spam, other, etc.

Another application of the proposed method is to support medical diagnosis of various diseases. Then, the main task of the algorithm is the induction of decision rules on the basis of medical research results from the database of many patients. The induced decision rules allow diagnosis of new patients. A typical example of a database and its analysis is the Breast Cancer Wisconsin Database (source: Dr. William H. Wolberg, University of Wisconsin Hospital, Madison, Wisconsin, USA) where the diagnosis of breast cancer for a new patient is supported by the database of nine attributes collected for 699 patients [25]. Another example is the analysis of the Pima Indians Diabetes Database of eight attributes and 768 female patients (source: National Institute of Diabetes and Digestive and Kidney Diseases, Maryland, USA), where the diagnostic binary-valued decision attribute investigates whether the patient shows signs of diabetes according to World Health Organization criteria.

Above examples demonstrate a crucial application of data mining and learning algorithms in the area of multimodal data processing. This is particularly evident in biomedical engineering where data collected for hundreds of variables describe multimodal medical parameters/measurements of patients. That implies a need to process large collections of data which include large databases of multimedia streams. For example, in [7] the classification of patients with the Alzheimer's disease was described. The research was based on brain Magnetic Resonance Imaging (MRI) processing and analysis.

In the area of data mining we deal with functional dependencies among attributes/parameters, i.e. we can make a research on that which attributes are necessary. For example, in form of decision tables we check which attributes could be removed without loss of any information. It is known as the feature extraction problem. Methods of feature extraction have proved very useful in many applications and have been studied by several researchers [1], [8], [9], [10], [13], [14], [15], [16], [17], [18], [20].

Noticeably, data mining and machine learning cannot exist without logic expressions. Such logic structure allows decision-making systems to learn from examples [2], [12], [20]. Therefore, the main goal of induction is the calculation of possibly shortest decision rules [19]. However, the compression is not the most important, but the fact that on the basis of such generalized knowledge one can make decisions for such data which is not included in the original decision system.

A significant difficulty in implementing decision-making systems is determined by efficient discretization of numeric data. For example, the attributes of the Pima Indians Diabetes Database include: number of times pregnant, plasma glucose concentration a 2 hours in an oral glucose tolerance test, diastolic blood pressure (mm Hg), triceps skin fold thickness (mm) 2-hour serum insulin (μU/ml) body mass index (weight in kg / (height in m)^2) diabetes pedigree function, age (years), and class variable (0 or 1). Most of these features are numeric, so for a proper analysis of this database it is necessary to discretize the data. A similar problem we face in the classification of electronic mails where records characterizing various network parameters used to perform anomaly analysis are often given as numeric values.

The problems (feature extraction, rule induction, data discretization) themselves cannot be polynomial, unless P = NP, because of its reducibility to CNF-satisfiability. However, use of some optimization techniques can make this process efficient for real-life examples.

2. Complementation algorithm

The most popular Algorithms for Discovering Rough Set Reducts methods are based on discernibility matrices [24]. In result these three tasks, i.e. feature extraction, rule induction, and data discretization, yields a discernibility function which is a Boolean formula in CNF. The simplification of the discernibility function is carried out by transforming the CNF into DNF. Such a trans-

formation is of NP computational complexity and therefore it is important to use efficient algorithms which can handle this task.

An interesting approach proposed by the author is based on the fast complementation algorithm [3], [4], [5]. The key strength of the algorithm lies in Shannon expansion procedure of monotone function *f*. Then,

$$f = \bar{x}_j f_{\bar{x}_j} + f_{x_j} \tag{1}$$

This procedure is fundamental in the field of logic synthesis, however it can successfully be applied in the field of data mining.

Proposed approach benefits from the transformation (2), i.e. double complementation of a Boolean function.

$$\prod_k \sum_l x_{kl} = \overline{\overline{\prod_k \sum_l x_{kl}}} = \overline{\sum_k \prod_l \bar{x}_{kl}} \tag{2}$$

Given that the discernibility function f_M representing the CNF is unate (monotone), it can be transformed into the F form (first complementation) and then considered as a binary matrix M (Fig. 1). In fact, the task of searching the complement of function F, i.e. $\bar{F}$, can be reduced to the concept of searching of a column cover C of the binary matrix M (second complementation).

Theorem [6]. Each row i of C, the binary matrix complement of M, corresponds to a column cover L of M, where $j \in L$, iff $C_{ij} = 1$.

The approach presented significantly accelerates calculations. An efficient representation of the algorithm in computational memory allows the authors to achieve results that cannot be calculated using published methods and systems. The performed experiments [3], [4], [5],[21], [22], [23], [25] have shown that despite many efforts directed to the designing of an effective tool for attribute reduction and data discretization, existing tools are not efficient.

Example.

Let's consider the discernibility function f_M as follows:

$$f_M = (x_2 + x_3 + x_4)(x_1 + x_2)(x_3 + x_4)(x_2 + x_3 + x_5).$$

Performing the multiplication and applying absorption law we obtain:

$$f_M = x_2x_3 + x_2x_4 + x_1x_3 + x_1x_4x_5.$$

The same result can be obtained performing the mentioned approach, i.e. double complementation of the function f_M. Then,

$$F = \bar{f}_M = \bar{x}_2\bar{x}_3\bar{x}_4 + \bar{x}_1\bar{x}_2 + \bar{x}_3\bar{x}_4 + \bar{x}_2\bar{x}_3\bar{x}_5,$$

and finally, applying Shannon expansion procedure, we calculate $\bar{F}$.

$$F = \begin{bmatrix} - & 0 & 0 & 0 & - \\ 0 & 0 & - & - & - \\ - & - & 0 & 0 & - \\ - & 0 & 0 & - & 0 \end{bmatrix} \rightarrow M = \begin{bmatrix} 0 & 1 & 1 & 1 & 0 \\ 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 \\ 0 & 1 & 1 & 0 & 1 \end{bmatrix}$$

$$\downarrow$$

$$\bar{F} = \begin{bmatrix} - & 1 & 1 & - & - \\ - & 1 & - & 1 & - \\ 1 & - & 1 & - & - \\ 1 & - & - & 1 & 1 \end{bmatrix} \leftarrow C = \begin{bmatrix} 0 & 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 1 \end{bmatrix}$$

Fig. 1. Diagram of the proposed algorithm

The illustrative diagram of the method has been shown in Fig. 1 and the full scheme of complementation of F using Shannon expansion in [3].

3. Bibliography

[1] Abdullah S., Golafshan L., Mohd Zakree Ahmad Nazri.: Re-heat Simulated Annealing Algorithm for Rough Set Attribute Reduction. In: International Journal of the Physical Sciences, vol. 6(8), pp. 2083–2089, 18 March, 2011.

[2] Bazan J., Nguyen Son H., Nguyen Sinh H., Synak P., and Wróblewski J.: Rough Set Algorithms in Classification Problems. In: Lech Polkowski, Tsau Young Lin, and Shusaku Tsumoto (eds.): Rough Set Methods and Applications: New Developments in Knowledge Discovery in Information Systems, volume 56 of Studies in Fuzziness and Soft Computing, pp. 49–88. Physica-Verlag, Heidelberg, Germany, 2000.

[3] Borowik, G., Łuba, T., Zydek, D.: Features Reduction using logic minimization techniques. In: Intl. Journal of Electronics and Telecommunications, vol. 58, No.1, pp. 71-76, (2012).

[4] Borowik G.: Data Mining Approach for Decision and Classification Systems Using Logic Synthesis Algorithms, in Advanced Methods and Applications in Computational Intelligence, s. Topics in Intelligent Engineering and Informatics, Eds. R.Klempous, J. Nikodem, W. Jacak, Z. Chaczko, Ch. 1, pp. 3-24, Springer Verlag 2013 (in print).

[5] Borowik G., Łuba T.: Fast Algorithm of Attribute Reduction Based on the Complementation of Boolean Function, in Advanced Methods and Applications in Computational Intelligence, s. Topics in Intelligent Engineering and Informatics, Eds. R.Klempous, J. Nikodem, W. Jacak, Z. Chaczko, Ch. 2, pp. 25-40, Springer Verlag 2013 (in print).

[6] Brayton R.K., Hachtel G.D., McMullen C.T., Sangiovanni-Vincentelli A.: Logic Minimization Algorithms for VLSI Synthesis, Kluwer Academic Publishers, 1984.

[7] Cuingnet R., Gerardin E., Tessieras J., Auzias G., Lehericy S., Habert M.O., Chupin M., Benali H., Colliot O.: Automatic classification of patients with Alzheimer's disease from structural MRI: A comparison of ten methods using the ADNI database. NeuroImage 56(2), 766–781 (2011).

[8] Dash R., Dash R., Mishra D.: A Hybridized Rough-PCA Approach of Attribute Reduction for High Dimensional Data Set. In: European Journal of Scientific Research, vol. 44(1), pp. 29–38, 2010.

[9] Feixiang Z., Yingjun Z., Li Z.: An Efficient Attribute Reduction in Decision Information Systems. International Conference on Computer Science and Software Engineering, pp. 466–469, Wuhan, Hubei, 2008.

[10] Gorska Z. and Janicki A.: Recognition of extraversion level based on handwriting and support vector machines, Perceptual and Motor Skills, vol. 3, no. 114, pp. 857–869, 2012.

[11] Grzenda, M.: Prediction-Oriented Dimensionality Reduction of Industrial Data Sets. In: Modern Approaches in Applied Intelligence, Mehrotra, K.G.; Mohan, C.K.; Oh, J.C.; Varshney, P.K.; Ali, M. (Eds.), LNAI 6703, 232–241 (2011)

[12] Grzymala-Busse J. W.: Data with Missing Attribute Values: Generalization of Indiscernibility Relation and Rule Induction. In: Peters J.F. et al. (eds.): Transactions on Rough Sets I, LNCS 3100, pp. 78–95, Springer-Verlag, Berlin, 2004.

[13] Jensen, R., Shen, Q.: Semantics-preserving dimensionality reduction: Rough and fuzzy-rough-based approaches. IEEE Transactions on Knowledge and Data Engineering, vol. 16, pp. 1457–1471, 2004.

[14] Jensen, R., Shen, Q.: Finding rough set reducts with ant colony optimization. Proceedings of the 2003 UK Workshop on Computational Intelligence, pp. 15–22, 2003.

[15] Janicki, A.; Staroszczyk, T.: Speaker Recognition from Coded Speech Using Support Vector Machines, In Proceedings of the 14th International Conference on Text, Speech and Dialogue (TSD 2011), Ivan Habernal and Václav Matoušek (Eds.). Lecture Notes on Artificial Intelligence (LNAI) 6836, pp. 291–298, Springer-Verlag, Berlin-Heidelberg, 2011

[16] Kryszkiewicz M., Cichoń K.: Towards Scalable Algorithms for Discovering Rough Set Reducts. In: Peters J.F. et al. (eds.): Transactions

on Rough Sets I, LNCS 3100, pp. 120–143, Springer-Verlag, Berlin 2004.
[17] Kryszkiewicz M., Lasek P.: FUN: Fast Discovery of Minimal Sets of Attributes Functionally Determining a Decision Attribute. In: Peters J.F. et al. (eds.): Transactions on Rough Sets IX, LNCS 5390, pp. 76–95, Springer-Verlag, Berlin, 2008.
[18] Nguyen D.T., Nguyen X.H.: A New Method to Attribute Reduction of Decision Systems with Covering Rough Sets. In: Georgian Electronic Scientific Journal: Computer Science and Telecommunications, vol. 1(24), pp. 24–31, 2010.
[19] Qinrong Feng, Duoqian Miao, Yi Cheng: Hierarchical decision rules mining, Expert Systems with Applications 37 2081–2091, 2010.
[20] Pawlak Z.: Rough Sets: Theoretical Aspects of Reasoning about Data. Kluwer Academic Publishers, 1991.
[21] ROSE2 – Rough Sets Data Explorer, http://idss.cs.put.poznan.pl/site/rose.html
[22] ROSETTA – A Rough Set Toolkit for Analysis of Data, http://www.lcb.uu.se/tools/rosetta/
[23] RSES – Rough Set Exploration System, http://logic.mimuw.edu.pl/~rses/
[24] Skowron A., Rauszer C.: The Discernibility Matrices and Functions in Information Systems. In: Słowiński R. (ed.): Intelligent Decision Support - Handbook of Application and Advances of the Rough Sets Theory, Kluwer Academic Publishers, 1992.
[25] UC Irvine Machine Learning Repository, http://archive.ics.uci.edu/ml/