DOI 10.2478/ama-2018-0033

# DATA MINING TECHNIQUES AS A TOOL IN NEUROLOGICAL DISORDERS DIAGNOSIS

**Małgorzata ZDRODOWSKA**[*], **Agnieszka DARDZIŃSKA**[*], **Monika CHORĄŻY**[**], **Alina KUŁAKOWSKA**[**]

[*]Faculty of Mechanical Engineering, Department of Biocybernetics and Biomedical Engineering,
Bialystok University of Technology, ul. Wiejska 45C, 15-351 Bialystok, Poland
[**]Faculty of Medicine, Department of Neurology, Medical University of Bialystok,
ul. M. Skłodowskiej-Curie 24A, 15-276 Białystok, Poland

m.zdrodowska@pb.edu.pl, a.dardzinska@pb.edu.pl, chorazym@op.pl, alakul@umb.edu.pl

*received 25 January 2018, revised 15 September 2018, accepted 19 September 2018*

**Abstract:** Neurological disorders are diseases of the brain, spine and the nerves that connect them. There are more than 600 diseases of the nervous system, such as epilepsy, Parkinson's disease, brain tumors, and stroke as well as less familiar ones such as multiple sclerosis or frontotemporal dementia. The increasing capabilities of neurotechnologies are generating massive volumes of complex data at a rapid pace. Evaluating and diagnosing disorders of the nervous system is a complicated and complex task. Many of the same or similar symptoms happen in different combinations among the different disorders. This paper provides a survey of developed selected data mining methods in the area of neurological diseases diagnosis. This review will help experts to gain an understanding of how data mining techniques can assist them in neurological diseases diagnosis and patients treatment.

**Keywords:** Data Mining, Classification Rules, Decision Tree, Action Rules, Neurological Disorders, Stroke, Multiple Sclerosis

## 1. INTRODUCTION

Neurological diseases are disorders that are associated with abnormal organic functioning of the peripheral and central nervous system. Neurological diseases are very dangerous and can even lead to death. They affect significantly the behavior of the patient and the functioning of his/her body. Neurological disorders can include the central nervous system, nerves and blood vessels that are designed to supply blood to the brain (Jacobs and Sapers, 2011; Kozubski and Liberski, 2003; World Health Organization, 2006).

In this article, we present data mining techniques for two of the most common types of neurological diseases:

- Stroke – sudden, local disturbance of blood circulation in the brain. Worldwide, it is the third leading cause of death and the main cause of disability of people over 40. There are two main types of strokes: hemorrhagic stroke and ischemic stroke. The main causes of strokes are: arterial hypertension, atrial fibrillation and ischemic heart disease, hypercholesterolemia, diabetes and others, e.g. alcoholism, smoking, obesity (Snarska et al., 2016; Trochimczyk et al., 2017; Yamashita, 2009).
- Multiple sclerosis (MS) – demyelinating disorder of the central nervous system (CNS), which includes the brain and the spinal cord. It is chronic inflammatory disease and immune-mediated disorder, which course in an individual patient is majorly unpredictable (Bejarano et al., 2011; Carreiro et al., 2011; Rodriguez et al., 2012). Multiple sclerosis can be associated with a variety of symptoms. One common symptom is difficulty of walking. Numbness or tingling in the hands or feet is another common symptom in the early stage of MS, which often goes undiagnosed as MS. Other commonly reported symptoms include: weakness or lost sensations in arms or legs, balance problems, poor coordination, tremor, difficulty articulating words, vision problems as well as bowel and bladder problems. Many of this symptoms can be caused by the other neurological diseases, that are not unique to MS, like extreme fatigue, low energy and becoming easily tired (Acquarelli et al., 2016; Ludwin et al., 2016).

## 2. SELECTED CLASSIFICATION METHODS

We work on data collected in form of information systems, which are defined as $S = (X, A, V)$, where:

- $X$ is a nonempty, finite set of objects;
- $A$ is a nonempty, finite set of attributes;
- $V = \{\cup V_a : a \in A\}$ is a set of all attributes values.

**Tab. 1.** Information system S

| $X$ | $a$ | $b$ | $c$ | $d$ |
|---|---|---|---|---|
| $x_1$ | $a_1$ | $b_2$ | $c_2$ | $d_1$ |
| $x_2$ | $a_1$ | $b_1$ | $c_1$ | $d_1$ |
| $x_3$ | $a_2$ | $b_1$ | $c_1$ | $d_1$ |
| $x_4$ | $a_2$ | $b_2$ | $c_1$ | $d_2$ |
| $x_5$ | $a_2$ | $b_2$ | $c_2$ | $d_2$ |
| $x_6$ | $a_2$ | $b_1$ | $c_1$ | $d_1$ |
| $x_7$ | $a_2$ | $b_2$ | $c_1$ | $d_2$ |
| $x_8$ | $a_2$ | $b_1$ | $c_2$ | $d_2$ |

Additionally, $a : X \to V_a$ is a function for any $a \in A$, that returns the value of the attribute of a given object. The attributes are divided into different categories: set of stable attributes $A_{St}$, set of flexible attributes $A_{Fl}$ and set of decision attributes $D$, such that $A = A_{St} \cup A_{Fl} \cup D.$ In this paper we analyze information systems with only one decision attribute $d$. The example of such a defined information system $S$ is represented as Tab. 1.

This information system is complete, and is represented by eight objects $\{x_i\}_{i=1,\dots,8}$ and four different attributes $\{a, b, c, d\}$. Attributes $\{a, b\}$ are stable (cannot be changed, e.g. gender, name), $\{c\}$ is flexible (can be changed, e.g. blood pressure), $\{d\}$ is a decision attribute (e.g. type of disease).

### 2.1. IF-THEN Rules

In data analysis the results of data mining have to be understandable for the user, so the analysis tools should provide clear results with the possibility of participating in the analysis process. An example of such tools are IF-THEN rules, which are a good technique of presenting information or part of knowledge (Han and Kamber, 2006; Lavrac and Zupan, 2010). An IF-THEN rule is as follows (Han and Kamber, 2006; Lavrac and Zupan, 2010):

IF *condition* THEN *conclusion*

For example, the rule R:

R: IF *gender* = *female*
AND *age* > 20
AND *difficulty walking* = *yes*
AND *numbness in the hands and feet* = *yes*
THEN *probability of MS* = *yes*

or

R: (*gender* = *female*) ∧ (*age* > 20) ∧ (*difficulty walking* = *yes*)
∧ (*numbness in the hands and feet* = *yes*)
⇒ (*probability of MS* = *yes*).

The IF part (left-hand side) is known as rule antecedent or precondition and the THEN part (right-hand) side is a rule consequent. In the rule antecedent, the conditional part can consist of more than one attribute tests, that are connected in a logical way, while the rule consequent contains only one class prediction. If, for a given tuple, all of the attribute tests in a rule antecedent are true, the rule antecedent is satisfied and the extracted rule covers the tuple (Han and Kamber, 2006).

All rules can be evaluated by the rule support and confidence. If the rule is in form $R: A \to B$, then:

$$sup\ (R) = card\ (A \cap B)$$

$$conf\ (R) = \frac{card\ (A \cap B)}{card\ (A)},$$

where the cardinality of a set is a measure of the number of elements of the set.

A higher value of the indicator indicates a better matched rule. It's worth mentioning that high accuracy on the training data, does not necessarily reflect true predictive accuracy. Many examples show, that rules supported by few examples have very high error rates (Lavrac and Zupan, 2010).

### 2.2. Decision Tree

A decision tree is an analytical decision support tool, a classifier in the form of a tree structure through which the decision can be made (Bejarano et al., 2013). The decision tree is similar to a flowchart, where each internal node indicates a test on an attribute, each branch means an outcome of a test and each leaf or terminal node is a class label. The root node is the highest node in a tree. The decision tree is very simple to visualize, understand, assimilate and interpret and it can manage high dimensional data of both types – numerical as well as categorical data. The user doesn't have to work a lot with data during preprocessing. He doesn't need to know any parameters settings or even the domain. Usually, decision tree classifiers are quite accurate, however, it depends on collected data (Han and Kamber, 2006; Larose, 2005; Pappa and Freitas, 2010; Triantaphyllou and Felici, 2006). A very simple example of decision tree classifier, which can help to predict multiple sclerosis is shown in Fig. 1. The predictors attributes are "*numbness in the hands and feet"* and *"difficulty walking"* and the classes are "*low probability of MS*" and "*high probability of MS*", indicating whether or not the patient may have multiple sclerosis.

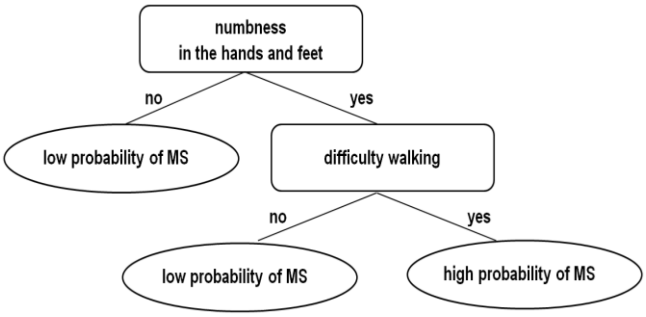

**Fig. 1.** An example of a decision tree

To classify any element by a decision tree, we go from the top of the tree, according to the branches, with attributes described in the element. We go down the tree until we reach a leaf node, which is one of the class (Pappa and Freitas, 2010). Analyzing the decision tree in Fig. 1, we can observe that the patient having the value *"numbness in the hands and feet*" = "*no*", would be assigned class "*low probability of MS*" regardless of the "*difficulty walking*" symptom, while the patient having the value "*numbness in the hands and feet*" = "*yes*" and "*difficulty walking*"="*yes*" would be assigned class "*high probability of MS*".

The most common algorithms of the decision trees are Iterative Dichotomiser 3 (ID3), which uses information increase as a measure of attribute selection and C4.5 (upgrade of ID3), which for attribute selection uses gain ratio. Because of its speed and accuracy, the algorithm C4.5 is used as a pattern for other classification algorithms. Collaterally with C4.5 algorithm, the Classification and Regression Tree (CART) algorithm was developed, which uses GINI index and makes binary decision trees. Classification trees are used when the dependent variable is categorical, while the regression trees are used when the dependent variable is continuous (Han and Kamber, 2006; Larose, 2005).

### 2.3. Action Rules

Action rules were first proposed by Raś and Wieczorkowska (2000) and are defined as a new class of rules, which includes directions for possible actions a user should take to reach a desirable point. An action rule is a rule obtained from an information system, that can characterize a transition, which may exist within objects from one state to another, concerning to user's decision attribute (Raś and Dardzińska, 2009; Raś and Wieczorkowska, 2000). Mining action rules is the process, which can identify some patterns in a decision system and shows the possibility of changes in object attributes. This knowledge may change the decision value (Dardzińska and Romaniuk, 2016; Dardzińska, 2013; Raś and Dardzińska, 2008a,b; Raś and Wieczorkowska, 2000). All in all, mining action rules work on the decision system with objects that have the following classes of attributes (Ludwin et al., 2016):

- Stable/semi-stable – attributes which value cannot be changed (or any change costs a lot). Consider the strokes data base, the examples of stable attributes can be age and sex.
- Flexible – attributes which value can be changed, e.g. relating to stroke, it can be arterial hypertension, hypercholesterolemia, diabetes, alcoholism, smoking, obesity, etc.
- Decisions – attributes where the user has a desire to change for more desirable state, e.g. type of therapy.

An example of action rules is shown in Tab. 1.

**Tab. 1.** An example of action rule

| | **stable attributes** | | | **flexible attributes** | | | | **decision** |
|---|---|---|---|---|---|---|---|---|
| patient | age | sex | … | alcoholism | arterial hypertension | diabetes | … | therapy |
| P1 | 65 | male | | no | yes | yes | | T1 |
| P2 | 70 | female | | no | yes | no | | T2 |
| P3 | 75 | male | | no | yes | no | | T0 |
| … | … | … | … | … | … | … | … | … |

**Action rule R:**
**(age > 65) ∧ (sex = male) ∧ (alcoholism = no) ∧ (diabetes, yes → no) ⇒ (therapy, T2 → T1)**

As we can see in Fig. 2 , an action rule can be represented in the following form:

$$[(\omega)\wedge(\alpha\to\beta)]\Longrightarrow(\Psi\to\Omega)$$

where $\omega$ indicates a fixed condition features conjunction, that is part of both groups, $(\alpha\to\beta)$ is recommended changes in flexible features value and $(\Psi\to\Omega)$ means an effect of the action, which the user wants to achieve. All action rules can be evaluated because of its support and confidence (Raś and Dardzińska, 2008, 2009).

## 3. APPLICATION IN NEUROLOGICAL DISORDERS

Classification methods in terms of the rules, decision trees, action rules, with definitions presented in Section 2 of this paper, are very interesting and promising in medical treatment fields. The knowledge can be extracted from a decision system that describes a possible transition of objects from one state to another with respect to a distinguished attribute called the decision attribute. They work on a set of classification rules extracted earlier. Certain pairs of these rules are combined to assign objects from one class to another. There is also a method which allows to explore action rules directly from the decision system. In Dardzińska (2013), the proposed algorithm, called Action Rules Discovery (ARD), builds rules for a given decision using an iterative marking strategy. It considers the change in attribute value as an atomic-action-term of length one, and then an action-term is a composition of atomic-action-terms. ARD starts by generating all atomic-action-terms for a given set of attribute values and assigning a mark (unmarked, positive, negative) based on standard support and confidence measures. The action-terms marked as positive are used to construct the action rules. The unmarked terms are placed into the list. From them all possible action-terms of length two are created. The process continues iteratively, creating terms of greater length, until the fixed point is reached.

Our dataset contains clinical data of 220 patients affected by strokes and 100 healthy patients. Patients are characterized by 55 attributes (20 stable and 35 flexible), and classified into three groups of strokes: ischemic stroke (IS), hemorrhagic stroke (HS) and transient ischemic attack (TIA). Ischemic stroke occurs as a result of an obstruction within a blood vessel supplying blood to the brain and affects for 87 percent of all stroke cases. Hemorrhagic stroke occurs when a weakened blood vessel ruptures. Transient ischemic attack is caused by a temporary clot. Often called a "mini stroke", these warning strokes should be taken very seriously. The goal was to find rules which help to reduce number of patients with strokes:

$$[class, IS\to healthy],$$
$$[class, HS\to healthy],$$
$$[class, TIA\to healthy].$$

We obtained several classification rules. Some of them are as follows:

- people who experience a TIA will go on to have a full-blown IS within a year (sup=33%, conf=95%);
- if a person has had a stroke, he/she will have another one within five years (sup=73%, conf=87%).

Seeking treatment reduces the chances of the disease. Therefore action rules were extracted. Some of them are given below:

– if patient with type IS is overweight and loses weight and begins regular physical activity, then his blood glucose and cholesterol returns to normal (sup=74%, conf=88%);
– if patient is a woman with type IS or HS is overweight and increases her physical activity, eating a healthy diet to maintain a normal weight and reduce drinking alcohol to no more than one per day then probability of having new stroke decreases rapidly (sup=82%, conf=78%);
– if patient is a man with type IS or HS is overweight and increases his physical activity, eating a healthy diet to maintain a normal weight and reduce drinking alcohol to no more than two per day for men, then probability of having new stroke decreases rapidly (sup=84%, conf=75%).

## 4. CONCLUSIONS

Data mining technology is being adopted in biomedical sciences and research for providing prognosis and deep understanding of the classification and verification of different disorders and diseases. Classifier systems in medical diagnosis are being developed progressively. Data mining techniques have been proposed to support the interpretation of medical data for clinical decision making, diagnosis or rehabilitation process. Most of these methods achieved promising prediction accuracies. But still researchers are working on different data sets, take different attributes into consideration, build their own feature selection models. Such knowledge makes difficult to propose a common comparison between methods. Therefore a road is for a unique, standardized method followed by imperative attribute selection and classification, after which we can obtain satisfying results.

## REFERENCES

1. **Acquarelli J., The Netherlands Brain Bank, Bianchini M., Marchiori E.** (2016), Discovering Potential Clinical Profiles of Multiple Sclerosis from Clinical and Pathological Free Text Data with Constrained Non-negative Matrix Factorization, In: Squillero G., Burelli P. (editors), Applications of Evolutionary Computation, *Lecture Notes in Computer Science*, Springer, Cham, 9597, 169–183.
2. **Bejarano H.B., Bianco M., Gonzalez-Moron D.** (2011), Computational classifiers for predicting the short-term course of Multiple Sclerosis, *BMC Neurology*, 11:67.
3. **Bejarano H.B., Segura V., Villoslada P.** (2013), Data mining in multiple sclerosis: computational classifiers. Introduction and methods (Part I), *Revista Española de Esclerosis Múltiple*, 5, 5–15.
4. **Carreiro A.V., Anunciação O., Carriço J.A., Madeira S.C.** (2011), Biclustering-Based Classification of Clinical Expression Time Series: A Case Study in Patients with Multiple Sclerosis, In: Rocha MP., Rodríguez JMC., Fdez-Riverola F, Valencia A. (editors), 5th International Conference on Practical Applications of Computational Biology & Bioinformatics, *Advances in Intelligent and Soft Computing*, Springer, Berlin, Heidelberg, 93.
5. **Dardzinska A.** (2013), *Action rules mining*, Springer-Verlag, Berlin.
6. **Dardzinska A., Romaniuk A.** (2016), Mining of Frequent Action Rules, In: Ryżko D, Gawrysiak P, Kryszkiewicz M, Rybiński H. (editors), *Machine Intelligence and Big Data in Industry*, *Studies in Big Data*, Springer, Cham, 19, 87-95.
7. **Han J., Kamber M.** (2006), *Data mining. Concepts and Techniques*, 2nd ed, Elsevier, San Francisco.
8. **Jacobs L.K., Sapers B.L.** (2011), Neurological Disease, In: Cohn S. (editor), *Perioperative Medicine*, Springer, London.
9. **Kozubski W., Liberski P.** (2003), *Neurological diseases* (in Polish), Wydawnictwo Lekarskie, Warsaw.
10. **Larose D.T.** (2005), *Discovering knowledge in data. An introduction to data mining*, John Wiley & Sons, Inc., New Jersey.
11. **Lavrač N., Zupan B.** (2010) Data Mining in Medicine, In: Maimon O., Rokach L. (editors), *Data Mining and Knowledge Discovery Handbook*, Springer, Boston.
12. **Ludwin S.K., Antel J., Arnold D.L.** (2016), Multiple Sclerosis, In: Pfaff D., Volkow N. (editors), *Neuroscience in the 21st Century*, Springer, New York.
13. **Pappa G.L., Freitas A.A.** (2010), *Automating the design of data mining algorithms. An evolutionary computation approach*, Springer – Verlag, Berlin.
14. **Raś Z.W., Dardzińska A.** (2008a), Action Rules Discovery Based on Tree Classifiers and Meta-actions, In: Rauch J., Raś Z.W., Berka P., Elomaa T. (editors), Foundations of Intelligent Systems, *Lecture Notes in Computer Science*, Springer, Berlin, Heidelberg, 5722.
15. **Raś Z.W., Dardzińska A.** (2008b), Action Rules Discovery without Pre-existing Classification Rules, In: Chan C.C., Grzymala-Busse J.W., Ziarko W.P. (editors), Rough Sets and Current Trends in Computing, *Lecture Notes in Computer Science*, Springer, Berlin, Heidelberg, 5306.
16. **Raś Z.W., Dardzinska A., Tsay L.-S., Wasyluk H.** (2008), Association Action Rules, *IEEE/ICDM Workshop on Mining Complex Data* (MCD 2008), 283–290.
17. **Raś Z.W., Wieczorkowska A.** (2000), Action-Rules: How to Increase Profit of a Company, In: Zighed D.A., Komorowski J., Żytkow J. (editors), Principles of Data Mining and Knowledge Discovery, *Lecture Notes in Computer Science*, Springer, Berlin, Heidelberg, 1910.
18. **Rodríguez J.P., Aritz P., Arteta D., Tejedor D., Lozano J.A.** (2012), Using Multi-Dimensional Bayesian Network Classifiers to Assist the Treatment of Multiple Sclerosis, *IEEE Transactions on Systems, Man, and Cybernetics-Part C: Applications and Reviews*, 42, 1705–1715.
19. **Snarska K.K., Bachórzewska-Gajewska K., Kapica-Topczewska K., Drozdowski W., Chorąży M., Kułakowska A., Małyszko J.** (2016), Hyperglycemia and diabetes have different impacts on outcome of ischemic and hemorrhagic stroke, *Archives of Medical Science*, 13(1), 100–108.
20. **Triantaphyllou E., Felici G.** (editors) (2006), *Data mining and knowledge discovery approaches based on rule induction techniques*, Springer Science+Business Media, New York.
21. **Trochimczyk A., Chorąży M., Snarska K.K.** (2017), An Analysis of Patient Quality of Life after Ischemic Stroke of the Brain, *The Journal of Neurological and Neurosurgical Nursing*, 6(2), 44–54.
22. **World Health Organization** (2006), *Neurological disorders: public health challenges*, Geneva.
23. **Yamashita T., Deguchi K., Sehara Y., Lukic-Panin V., Zhang H., Kamiya T., Abe K.** (2009), Therapeutic strategy for ischemic stroke, *Neurochemical Research*, 34, 707–710.

Acknowledgements: This work is supported by the Ministry of Science and Higher Education of Poland under research project No. MB/WM/17/2018.