DECISION-MAKING PROCESS IN COLON DISEASE AND CROHN'S DISEASE TREATMENT

Agnieszka DARDZINSKA*, Anna KASPERCZUK*

*Faculty of Mechanical Engineering, Department of Biocybernetics and Biomedical Engineering,
Bialystok Technical University, ul. Wiejska 45C, 15-351 Bialystok, Poland

a.dardzinska@pb.edu.pl, a.kasperczuk@ pb.edu.pl

Abstract: The article presents the process of building a logistic regression model, which aims to support the decision-making process in medicine. Currently, there is no precise diagnosis for ulcerative colitis (UC) and Crohn's disease (CD). Specialist physicians must exclude many other diseases occurring in the colon. The first goal of this study is a retrospective analysis of medical data of patients hospitalized in the Department of Gastroenterology and Internal Diseases and finding the symptoms differentiating the two analyzed diseases. The second goal is to build a system that clearly points to UC or CD, which shortens the time of diagnosis and facilitates the treatment of patients. The work focuses on building a model that can be the basis for the construction of classifiers, which are one of the basic elements in the medical recommendation system. The developed logistic regression model will define the probability of the disease and will indicate the statistically significant changes that affect the onset of the disease. The value of probability will be one of the main reasons for the decision.

Key words: Colon Disease, Logistic Regression, Decision System, Information System

1. INTRODUCTION

Let us assume that $S = (X, A, V)$ is an information system, where (Dardzinska, 2013; Kasperczuk and Dardzinska, 2016):

- X is a nonempty, finite set of objects;
- A is a nonempty, finite set of attributes;
- V is a set of all attributes value.

Then, $a : X \rightarrow V_a$ is a function for any $a \in A$, that returns the value of the attribute of a given object. The attributes are divided into three categories: attributes A_1, flexible attributes A_2 and decision attributes D (in A_2 and D the values of attributes can change), such that $A = A_1 \cup A_2 \cup D$ (Han and Kamber, 2006; Pauk and Dardzinska, 2012, Dardzinska and Romaniuk, 2016; Gurdal and Dardzinska, 2017).

Tab. 1. Information System

Object	STab. attributes A_1	Flexible attributes A_2	
	Attribute a	Attribute b	Attribute c
x_1	$a1$	$b1$	H
x_2	$a1$	$b2$	H
x_3	$a1$	$b3$	I
x_4	$a2$	$b3$	I
x_5	$a2$	$b2$	H

Example of the information system $S = (X, A, V)$ is presented in Tab.1. The set of objects consists of five elements $X = \{x_1, x_2, x_3, x_4, x_5\}$. The set of attributes consists of two subsets (Agrawal and Srikant, 1993; Dardzinska and Ras, 2011):

- A_1 which includes sTab. attributes $\{a\}$;
- A_2 which includes only flexible attributes $\{b, c\}$.

Information systems can be also seen as decision table. In Tab. 2 we have decision System $S = (X, A, V \cup \{d\})$, with one sTab. attribute a, two flexible attributes b and c and the decision attribute d.

Tab. 2. Decision System

Object	Atribute a	Attribute b	Attribute c	Decision d
x_1	$a1$	$b1$	H	-
x_2	$a1$	$b2$	H	-
x_3	$a1$	$b3$	I	+
x_4	$a2$	$b3$	I	-
x_5	$a2$	$b2$	H	+

2. CHRONIC ENTERITIS

Chronic intestinal inflammations have probably afflicted mankind since the dawn of time. As soon as science, thanks to the work of Antoni Leeuwenhoek in the eighteenth century, and then Ludwik Pasteur in 100 years later, discovered the world of microorganisms and proved their relationship with many diseases, people began to look for the causes of these ailments in infections. This trend also exists today (de Kruif, 1956). Nevertheless, there were researchers who described morphological changes in the intestinal wall involved in the disease process. These changes were characteristic of persistent inflammatory conditions occurring in the gastrointestinal tract, mainly in its distal part. Already in 1761, the Italian anatomist Giovanni Battista Morgagni described the case of a 34-year-old man with inflammatory infiltrations of the small intestine and large intestine, as well as ulcers, stenoses and

enlarged mesenteric lymph nodes. Other authors such as Samuel Wilks also described similar lesions, and in 1859 he introduced the term ulcerative colitis (UC) in the nomenclature, which describes the cases of severe ulcerations in the colon, in patients with profuse, bloody diarrhea (Kirsner, 1988).

In the medical literature, the term chronic inflammatory bowel disease, also re-ferred to as Inflammatory Bowel Disease (IBD), has been functioning since 1932, when New York doctor Burrill B. Crohn and his colleagues Leon Ginzburg and Gor-don D. Oppenheimer publish a work in which they present a new disease unit called "regional ileitis" - from this point medicine knows the disease as Crohn's disease (CD) (Crohn et al., 1932; Aufses, 2000).

Long before Crohn's publication of his work, doctor Antoni Lesniowski (Warsaw) repeatedly describes cases of chronic inflammatory bowel states, with marked in-flammatory changes of the intestinal mucosa and enlarged local lymph nodes as well as fistulas (Lesniowski, 1903; Lichtarowicz and Mayberry, 1988).

2.1. Crohn's Disease

Crohn's disease is a chronic inflammatory disease that can affect every part of the digestive tract, from the mouth to the anus. Manifested symptoms depend on where the inflammatory changes are located. Generally, it can be distinguished: abdominal pain (usually in the right iliac fossa), bloating, weight loss, weakness and fever. It is also observed the formation of perianal fistulas and abscesses. Inflammatory changes in the gastrointestinal tract are discontinuous, which means that next to the inflamed mucous membrane, it is unchanged inflammation. In addition, inflammatory changes in addition to the mucous membrane also include all the other parts of the intestinal wall (Hauser et al., 2007).

2.2. Ulcerative colitis

Like CD, it is a chronic inflammatory disease, affects only the colon. Symptoms include frequent diarrhea with an admixture of blood and / or mucus and painful pressure on the stool. Fistulas and abscesses are rare. Permanent inflammatory changes are localized along the entire length of the large intestine in the form of superficial ulcerations. The histological image shows granulocytes and lymphocytic infiltrates of the mucous membrane. At an advanced stage, i.e. after about 10 years of the disease, epithelial dysplasia changes leading to tumor degeneration may appear (Hauser et al., 2007).

3. LOGISTIC MODEL

Logistic regression is a frequently used statistical method for classification problems when the variable is explained on a dichotomous scale. This means that the predictive model of logistic regression determines the probability of one of two possible outcomes: pathologies or no pathologies. Logistic regression is based on a specific way of expressing probability, called the odds ratio (OR). The opportunity quotient determines the probability as the ratio of probability of success to probability of failure.

$$S(A) = \frac{p(A)}{p(\neg A)} = \frac{p(A)}{1-p(A)}. \tag{1}$$

It can be calculated using the above formula. The main advantage of the odds ratio, as compared to conventional probability is that the odds ratio assumes values in the range $(0, +\infty)$ for p range from 0 to 1, and the logarithm value of the field $(-\infty, +\infty)$. This means that we can use regression methods not limited to a range (0-1) such as linear regression to estimate the log of chance in a regression model. S (A) determines the odds ratio for malignancy in patients with known risk factors for cancer for pathological survival in patients who do not. For tumour diagnostics, the odds ratio greater than 1.0 indicates that the risk of hyperplasia or altered cells in patients with increased risk of OR is increased, while S (A) <1.0 allows for exclusion of pathology (Hosmer et al 2013).

Definition 1. _Odds ratio_

For the odds ratio, a 95% confidence interval is given, the span is based on the number of patients in the study group. The odds ratio can also be calculated taking into account the division of the respondents into two separate groups using the following formula:

$$OR_{AxB} = \frac{S(A)}{S(B)} = \frac{p(A)}{1-p(A)} \div \frac{p(B)}{1-p(B)}. \tag{2}$$

We interpret them as follows:
- If OR> 1, then in the first group the occurrence of the event is more likely.
- If OR <1, the event occurs in the second group more likely.
- If OR = 1, then the event is equally likely in both classes.

Definition 2. _Likelihood function_

We derive the form of the reliability function L for logistic regression. The explanatory variable Y is binary and for single observation i and occurs:

$$Y_i|X_i = \begin{cases} 1 \ with \ probability \ p(X_i) \\ 0 \ with \ probability \ 1 - p(X_i) \end{cases}, \tag{3}$$

$$L(X_i, \beta) = P(Y_i = 1|X_i)^{Y_i} \cdot P(Y_i = 0|X_i)^{1-Y_i} = p(X_i)^{Y_i}[p(X_i)]^{1-Y_i}. \tag{4}$$

Definition 3. _Logistic function_

Transformation function on the logarithm of the probability of chance is called logit:

$$logit(P) = ln\frac{p}{1-p} = \ln(p) - \ln(1-p), where \ p = \frac{e^{logit(p)}}{1+e^{logit(p)}} = \frac{1}{1+e^{logit(p)}}. \tag{5}$$

The logistic model is based on the function $f(z)$:

$$f(z) = \frac{e^z}{1+e^z} = \frac{1}{1+e^{-z}}. \tag{6}$$

The predictive model predicts the probability of malignancy with the probability of benign tumours. The constructed model also allows us to observe which of the independent variables tested can influence the dependent variable explained on a dichotomous scale. Conditional probability that the dependent variable Y assumes a value of 1 for the value of the independent variables $x_0, x_1, ..., x_k$ can be described using a following pattern:

$$P(Y = 1|x_0, x_1, ..., x_k) = \frac{e^{(a_0 + \Sigma_{i=1}^k a_i x_i)}}{1+e^{(a_0 + \Sigma_{i=1}^k a_i x_i)}}, \tag{7}$$

where $a_i, i = 0 \dots k$ are regression coefficients, while $x_1, x_2, \dots, x_3$ represent independent variables. The independent variable in logistic regression model can be quantitative or qualitative. Values of estimators are calculated using the most reliable method. The greater reliability of a model, the more likely it is that the variable will appear in the sample, which means better matching model to the data (Jong and Heller, 2008; Harrell, 2001).

Definition 4. *Confusion matrix*

The error Tab. 3 (error matrix) is used to assess the quality of a binary classification (into two classes). Data marked with positive and negative labels are subjected to a classification that assigns them a predicted positive class or a predicted negative class. It is possible that the data originally marked as positive will be mistakenly classified as negative. All such a situation is presented in the Tab. 3 of errors (Jong and Heller, 2008).

The board has two rows and two columns. Rows represent predicted classes, while columns represent real classes.

Tab. 3. Confusion matrix

Observed event	Expected event	
	Number 1	**Number 2**
Number 1	TP (true positive)	FN (false negative)
Number 2	FP (false positive)	TN (true negative)

Definition 5. *kNN-algorithm*

In order to improve the performance of the k-NN algorithm, the commonly used technique is standardization or normalization of data. Its use causes all dimensions for which the distance is calculated to have equal significance. Otherwise, one dimension would be dominated by other dimensions.

Standardization consists in bringing a situation where the average value of a particular feature is 0 and the standard deviation is equal to 1,

$$x_j(i) = \frac{x_j(i) - mean(x_j)}{std(x_j)}, \tag{8}$$

where: i – another vector index, j – index of feature (variable), $mean(x_j)$ – average value of variable j, $std(x_j)$ – standard deviation of variable j.

Normalization consists in bringing about a situation where the value of the variable belongs to the range [0,1]. It is expressed by the formula:

$$x_j(i) = \frac{x_j(i) - min(x_j)}{max(x_j) - min(x_j)}, \tag{9}$$

where: i – another vector index, j – index of feature (variable), $max(x_j)$ – maximum value of variable j, $min(x_j)$ – minimum value of variable j.

The description of k-NN algorithm, consisting of teaching and testing modules [Liu et. al., 2003] which looks as follows:
Teaching:
1. Make alternatives: standardize / normalize / leave data as they are.
2. Remember the entire training set.
Testing:
1. Make standardization /normalization/ leave the data that are (testing).
2. Count the distance between the test vectors and all the vectors of the training set.

3. Sort the distances from the largest to the smallest.
4. See the k-label of the closest vectors to the test vector. Make a histogram of the frequency of each label from "k-nearest" (how many and whose labels were among the nearest ones).
5. Assign the most common label as a test vector label.
6. If there was an impasse (two classes had the same number of votes) solve the problem randomly.

Model Testing

We use some statistics, which are explained briefly as follows (Hosmer et al. 2013; Olson and Durson, 2008; Powers, 2011):
− TP Rate, recall or sensitivity – rate of the instances correctly classified as a given class). Proportion of instances classified as a given class divided by the actual total in that class. Describes the ability to detect really ill people (who have a given feature). So, if we are examining a group of people who are sick, then sensitivity gives us information as to what percentage of them has a positive test result):

$$TPR = \frac{TP}{TP + FN}. \tag{10}$$

− Specificity – the ability to detect people who are actually healthy (without a given trait). For example, when we study a group of healthy people, then the specificity gives us information on the percentage of them that have a negative test result:

$$TNR = \frac{TN}{TN + FP}. \tag{11}$$

− FP Rate - rate of the instances falsely classified as a given class:

$$FPR = \frac{FP}{FP + TN}. \tag{12}$$

− Precision – proportion of instances that are truly of a class divided by the total instances classified as that classes:

$$PPV = \frac{TP}{TP + FP}. \tag{13}$$

− F-Measure - indicator of quality of the model, the harmonic mean of precision and sensitivity:

$$F = 2 \cdot \frac{PPV \cdot TPR}{PPV + TPR}. \tag{14}$$

− ROC Curve – the area under the ROC Curve. The accuracy of the test depends on how well the test separates the group being tested into those with and without the disease in question.
− Kappa Statistic – a measure of conformity between the proposed allocation instance of the class and the actual, which is about the overall accuracy of the model.
− Number of correctly/incorrectly classified instances.

4. RESULTS

In the experiment, we use the data of patients suffering from ulcerative colitis (86 cases) and Crohn's disease (66 cases). The total distribution is shown in Tab. 4.

The analysis was based on the construction of two logistic models with various attributes. In the first stage of the analysis, we examined the significance and the model that takes into account the attributes that specialists pay more attention to the diagnosis. Attributes are as follows: WBC (numeric), RBC (numeric),

Hg (numeric), platelets (numeric), albumin (numeric), fibrinogen (numeric), abdominal tumor (nominal {0, 1}) and CRP (numeric).

The codes of classes:

$$Y = \begin{cases} 0 \ WZJG \\ 1 \ ChLC \end{cases}. \tag{15}$$

Tab. 4. Distribution of groups

Class	Men	Women	Totality
0 (UC)	54	32	86
1 (CD)	34	32	66
Totality	88	64	152

Tab. 5. Coefficients and OR for model 1

Attribute	Coefficient (Class 0)	OR
WBC(white blood cel)	0.0204	1.0206
RBC(red blood cel)	-1.9254	0.1458
Hemoglobin (Hg)	0.5623	1.7547
Platelets (PLT)	0.0007	1.0007
Albumin	-0.2601	0.771
Fibrynogen	-0.0019	0.9981
Abdominal_tumor	17.5061	40.9792
CRP	0.0086	1.0087
Intercept	2.8915	

Tab. 6. Confusion matrix for model 1

Observed event	Expected event	
	UC	CD
UC	65	21
CD	35	31

$$sensitivity = \frac{TP}{TP+FN} = \frac{60}{100} = 65\% , \tag{16}$$

$$specificity = \frac{TN}{FP+TN} = \frac{31}{52} = 60\% . \tag{17}$$

The obtained results indicate that almost all selected attributes have no significant impact on the classification of patients to the disease entity. The classifier often assigns the patient randomly to a given group. The only attribute that really distinguishes a patient with ulcerative colitis and Crohn's disease is the presence of a tumor in the abdomen (OR $= 40.9792, 95\%$ CI: $39.9871 \div 42.5643$). This indicates that patients who have been diagnosed with cancer suffer from ulcerative colon cancer (class 0). In the case of RBC, the OR is 0.1458 (95% CI: $0.0981 \div 0.1998$), which means that an increase in its value should cause the patient will be classified in the group with Crohn's disease. For the remaining attributes, OR assumes values oscillating around the value 1. It can be assumed that their values do not differ in both groups of patients.

In addition, we examined whether attributes significantly affect the decision attribute. Significance tests were conducted at the level of significance $\alpha = 0,05$. Only attribute WBC is statistically significant (Z $= -2.01744$; p $= 0.04365$). The OR for WBC is equal to 1.02. The other attributes did not significantly influence the decision regarding belonging to a given group of patients. The results indicate that the currently used classification model of the patient is not always clear. It may introduce some discrepancies,

especially if the diagnosis is being attempted by a young, less experienced specialist.

For these reasons, we decided to use statistical and machine learning methods in order to build a new and an improved model, which could more accurately classify patients into two groups: patients with UC and CD.

We conducted the selection process using the Correlation-based Feature Subset Selection for Machine Learning, which limited the collection to 7 attributes. It evaluates the value of a subset of attributes, taking into account the individual predictive ability of each attribute along with the degree of redundancy between them. Preferred ones are subsets of features that are strongly correlated with the class at low intercorrelation. The second model contains following attributes: age (numeric), smoke ({0, 1}), blood feces ({0, 1}), eosinophils (numeric), alanine aminotransferase (numeric), sodium (numeric), potassium (numeric).

In the case of the proposed model (Tab. 7), OR values that significantly deviate from the value 1 were obtained for the smoke attribute (OR $= 0.0118, 95\%$ CI: $0.0001 \div 0.0234$). This means that people who smoke more often suffer from CD. We can conclude that smoking does not affect the development of UC. Another attribute is blood feces. In this case OR $= 9.6569, 95\%$CI: $9.9872 \div 10.9872$, which means that the phenomenon of bloody stools much more often means UC than CD. The occurrence of blood stools about nine times increases the probability of ulcerative colitis.

Tab. 7. Coefficients and OR for model 2

Attribute	Coefficient (class 0)	OR
Age	0.055	1.0565
Smoke	-4.442	0.0118
Blood feces	2.2677	9.6569
Eosinophils	-1.0018	0.3672
alanine aminotransferase	0.0465	1.0475
Sodium	0.2827	1.3267
Potassium	-2.0809	0.1248
Intercept	-31.4691	

Tab. 8. Confusion matrix for model 2

Observed event	Expected event	
	UC	CD
UC	74	12
CD	16	50

$$sensitivity = \frac{TP}{TP+FN} = \frac{74}{100} = 74\% , \tag{18}$$

$$specificity = \frac{TN}{FP+TN} = \frac{50}{52} = 96\% . \tag{19}$$

Analysing the OR values, we can conclude that an increase in the sodium value by one results in a 30% increase in the certainty that a given patient belongs to the UC group. Higher values of potassium and eosinophils are indicative of Crohn's disease.

In addition, analysis of significance indicates that all parameters of the model parameter impact on decision-making at the level of significance $\alpha = 0.05$.

In order to compare the quality of constructed models. measures were calculated. All-important measures are presented in Tab. 9. This indicate that a model 2 classifies patients better than model 1. All quoted measures assume values in favour of the second model. This is particularly evident in the case of class 1 (Crohn's disease). The sensitivity and specificity for second model have higher value. This is particularly evident in the case of the specificity value, which is 96% and indicates a very good quality of the predictive model.

Tab. 9. Measure comparing models

Class		TP Rate	FP Rate	Precision	F-Measure	AUC
0	Model 1	0.802	0.545	0.657	0.723	0.640
	Model 2	0.86	0.242	0.822	0.841	0.921
1	Model 1	0.455	0.198	0.638	0.531	0.640
	Model 2	0.758	0.14	0.806	0.781	0.921

Model 1 correctly classified 65% of cases, with the proposed model 2 correctly classifying 92% of cases. This proves the very good quality of the constructed classifier. In addition, the values of Kapp's statistics were compared. In the case of the first model it amounted to 27% and for the second model it was 82%.

5. CONCLUSIONS

Storing a huge amount of data does not bring measurable benefits - it is necessary to analyze them so that you can get potentially important knowledge hidden in them. Proper use of the possessed data gives a unique opportunity to find the most relevant information and dependencies. In many cases classification methods are very useful in modern medicine. They are very helpful in finding new symptoms and patients treatment methods. In this paper, we built classification model for dependent variable. It becomes an important tool to find symptoms that affect whether the patient is ill or not. The proposed classification model correctly classifies new cases (correctly classified instances – 92%, Kappa Statistics – 82%). With the help of the built-in classifier, it will be possible to automatically diagnose subsequent patients.

REFERENCES

1. **Agrawal R., Srikant R.** (1993), Fast algorithm for mining assocation rules, *International Conference on Very Large Databases*, 487–499.
2. **Aufses A.H.** (2000), The History of Surgery for Crohn`s Disease at The Mount Sinai Hospital, Mount Sinai Journal of Medicine, 67, 198–203.
3. **Crohn B.B., Ginzburg L., Oppenheimer G.D.** (1932), Regional ileitis. A pathologic and clinical entity, Journal of the American Medical Directors Association, 99, 1323–1329.
4. **Dardzinska A.** (2013), *Action rules mining*, Springer-Verlag, Berlin.
5. **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.
6. **De Kruif P.** (1956), Microbe hunters (In Polish: *Łowcy mikrobów)*, Państwowy Zakład Wydawnictw Lekarskich, Warszawa.
7. **Gürdal O., Dardzinska A.** (2017), A New Approach to Clinical Medicine by Action Rules, International Journal of Development Research, 7(1), 11032–11039.
8. **Han J., Kamber M.** (2006), *Data Mining: Concepts and Techniques*, Morgan Kaufmann Publishers, Second Edition, 21–27.
9. **Harrell F.** (2001*), Regression Modeling Strategies with Applications to Linear Models*, Logistic Regression, and Survival Analysis, Springer-Verlag, New York.
10. **Hauser W., Hoffmn J., Kuhbacher T., Raible A., Reinshagen M., Rogler G., Schreiber S., Neukamm U., Eceterski A.** (2007), Crohn's disease and other inflammatory bowel diseases (In Polish:*Choroba Lesniowskiego - Crohna i inne nieswoiste zapalenia jelit)*, Wydawnictwo Lekarskie PZWL.
11. **Hosmer D., Lemeshow S., Sturdivant R.** (2013), Applied Logistic Regression, Wiley Series in Probability and Statistics.
12. **Jong de P., Heller G.Z.,** (2008), *Generalized Linear Models for Insurance Data*, Cambridge University Press, Cambridge.
13. **Kasperczuk A., Dardzinska A.,** (2016), Comparative Evaluation of the Different Data Mining Techniques Used for the Medical Database, *Acta Mechanica et Automatica*, 10(3), 233–238.
14. **Kirsner J.B.** (1988), Historical aspects of inflammatory bowel disease, *Journal of Clinical Gastroenterology* ,10, 286–297.
15. **Lesniowski A.** (1903), A contribution to the surgery of the bowels (In Polish:Przyczynek do chirurgii kiszek), *Medycyna*, 31, 460–518.
16. **Lichtarowicz A.M., Mayberry J.F.** 1988, Antoni Lesniowski and his contribution to regional enteritis (Crohn`s disease), *Journal of the Royal Society of Medicine*, 81, 468–470.
17. **Liu T., Moore A., Gray A.,** (2003) Efficient Exact k-NN and Nonparametric Classification in High Dimensions, *Advances in Neural Information Processing Systems*, 16, 8–13.
18. **Olson D., Dursun D.** (2008), *Advanced Data Mining Techniques*, Springer.
19. **Pauk J., Dardzinska A.** (2012), New method for finding rules in incomplete information systems controlled by reducts in flat feet treatment, Image Proc. and Communications Challenges, *Advances in Intelligent and Soft Computing*, 184, 209–214.
20. **Powers D.** (2011), Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness & Correlation, *Journal of Machine Learning Technologies*, 2(1), 37–63.
21. **Ras Z., Dardzinska A.** (2011), From Data to Classification Rules and Action, *International Journal of Intelligent Systems*, Wiley, 26(6), 572–590.