VERSITA
Information Technology and Management Science
doi: 10.2478/v10313-012-0018-3
2012 /15

The Use of BEXA Family Algorithms in Bioinformatics Data Classification

Madara Gasparovica[1], Ludmila Aleksejeva[2], Valdis Gersons[3], [1-3]*Riga Technical University*

***Abstract* – This article studies the possibilities of BEXA family classification algorithms – BEXA, FuzzyBexa and FuzzyBexa II in data, especially bioinformatics data, classification. Three different types of data sets have been used in the study – data sets often used in the literature, UCI data repository real life data sets and real bioinformatics data sets that have the specific character – a large number of attributes and a small number of records. For the comparison of classification results experiments have been carried out using all data sets and other classification algorithms. As a result, conclusions have been drawn and recommendations given about the use of each algorithm of BEXA family for classification of various real data, as well as an answer has been given to the question, whether the use of these algorithms is recommended for bioinformatics data.**

***Keywords* – classification algorithms, bioinformatics data, BEXA, UCI data**

I. Introduction

This article studies the possibilities of Bexa type algorithms in real bioinformatics data classification. Bexa family of algorithms consists of three separate algorithms: Bexa that works with crisp data, FuzzyBexa and FuzzyBexa II that both work with fuzzy data.

Bexa (Basic EXclusion Algorithm) algorithm was developed in 1996 by Theron and Cloete [1]. BEXA is a covering algorithm, which belongs to the classical (crisp) inductive learning classification algorithm group. Its working is based on the use of exclusion principle in the process of inductive learning. To learn more about this algorithm, see Section II.

FuzzyBexa algorithm was developed in 2004 by Zyl and Cloete [2]. The task of FuzzyBEXA algorithm is to create a good set of rules for further classification. The process of classification itself is not among the tasks of this algorithm, but, for more clarity on this issue, this subsection pays attention to the execution of classification process. Usually all instances are not covered by a created rule. In this case, FuzzyBEXA algorithm creates a default rule, which covers all possible instances. For more information about this algorithm, see Section II.

Algorithm FuzzyBexaII was also developed in 2004 by modifying FuzzyBexa algorithm [3]; it was created by the authors of FuzzyBexa - Zyl and Cloete. In this algorithm each class is not examined individually; instead, it generates rules for all classes. For more information about algorithm see Section II.

The experiments in this paper are carried out using sixteen real data sets that can be conditionally divided into three parts: data sets often used in the literature (Iris data set, Auto MPG and Ionosphere data set), UCI data repository real life data sets (Nursery data set, Breast cancer Wisconsin, Parkinson, SPECT heart, Molecular biology (Splice-junction gene sequences), Yeast dataset) and real bioinformatics data sets that have the specific character – a large number of attributes (several thousands) and a small number of records (GSE3726 (Breast & colon cancer), GSE2535 (CML treatment), GSE2685 (Gastric cancer), GSE1577 (Lymphoma & Leukaemia), GSE2191 (AML prognosis), GSE89 (Bladder cancer) and GSE1987 (Lung cancer)). For more information about the data sets see Section III.

For the comparison of classification results, experiments have been carried out using all data sets and other classification algorithms. Bexa is compared to JRip, Part and PRISMA algorithms for categorical data, as well as JRip and Part for continuous data. FuzzyBexa and FuzzyBexa II are compared to FURIA, FLR and Slave C algorithms. The classification results are given in Section IV.

As a result, conclusions have been drawn and recommendations given about the use of each algorithm of BEXA family for classification of various real data sets, as well as an answer has been given to the question, whether the use of these algorithms is recommended for bioinformatics data. For more information see Section V.

From the obtained classification results it can be seen that the use of algorithms from Bexa family in bioinformatics is perspective, and more research is needed to improve the deficiencies of the algorithms to increase their classification accuracy and the quality of obtained rules.

II. The Used Algorithms

This Section describes all three of the used algorithms: BEXA, FuzzyBEXA and FuzzyBEXA II. The overall working scheme of algorithm execution is showed in Fig. 1.

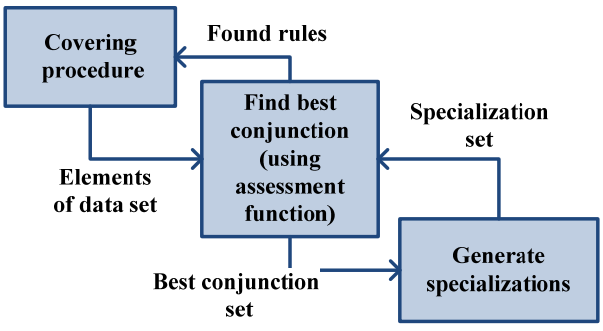

Fig. 1. BEXA family algorithm scheme

BEXA is a covering algorithm that belongs to the classical (crisp) group of inductive learning classification algorithms.

Its functionality is based on the use of exclusion principle in the process of inductive learning. A significant feature of BEXA algorithm is that it uses a general-to-specific searching principle. In the case of general-to-specific searching principle BEXA algorithm starts with a general conjunction (rule) and proceeds with concretization of the conjunction. This process is executed until one of the stopping criteria (stop growth test) is reached. The main task of the general-to-specific searching principle is related to the necessary decrease in number of the specializations and increase in quality in each step of the algorithm. Too many specializations can lead to inadequately high use of resources; on the other hand, too few specializations decrease the probability to find the best conjunctions. Some algorithms tackle this problem by using strict and inelastic conditions, e.g., CN2 algorithm constructs only 'clean' conjunctions [4], but AQ15 algorithm maximally considers only as many specializations as there are attributes in the data set [5]. The use of such strict conditions has a significant drawback – there is a chance to miss and not build potentially good specializations. In the case of BEXA algorithm, the conditions are dynamic, which allows using the overall features of a data set in the selection process of specializations [1]. BEXA algorithm dynamic conditions position it as a suitable algorithm for creation of precise and simple conjunctions.

Usually BEXA algorithm is divided into three main procedures [1]:

(1) Covering procedure (COVER-P);

(2) Procedure of finding the best conjunction (Find-Best-Conjunction);

(3) Procedure of creating specializations (Generate-Specializations).

A. FuzzyBEXA

The structure of fuzzy data classification algorithm FuzzyBEXA is based on crisp data classification algorithm BEXA [6]. FuzzyBEXA algorithm expands the use of definitions described in BEXA algorithm to their application to fuzzy data. In the case of the algorithm of classical data classification BEXA, the set of conjunction covered instances is considered to be all records that fit the given conjunction [6]. In this case, a clearly defined value of a specific attribute either fits or does not fit the conjunction. In the case of fuzzy data classification algorithm FuzzyBEXA, the value of an attribute fits the conjunction in the scale from 0 to 1, and therefore a record can fit the conjunction with a very small membership indicator. Such situation may be undesirable; therefore, new variables are used "*alpha-cut*" and "*alpha-class cut*" [6].

1. Variable "Alpha-Cut"

Variable "*alpha-cut*" (or *alpha-leveling*) (α_a) determines that all membership values of a record that are below the level of this variable value are considered 0 [33]. Thus, the instance set covered by a conjunction can be defined as follows (see Equation (1)):

$$X_s(\mathrm{c}) = \{\mathrm{s} \in \mathrm{S} \mid \mu_{\mathrm{c}}(\mathrm{S}) \geq \alpha_a\} \qquad (1)$$

where X_s (c) – the record set covered by conjunction,

s – the record,

S – the record set,

$\mu_{\mathrm{c}}(\mathrm{S})$ – the membership function of a record for attribute a ,

c - the conjunction,

α_a – the alpha-cut variable for attribute a,

a – the attribute identifier.

2. Variable "Alpha-Class Cut"

For BEXA tree algorithms to function correctly, there is a necessity to divide the data into positive and negative class records. The problem is that such division in the case of fuzzy data is not directly possible. It is explained by the fact that values of each record, which are similar to attributes, and the class of a record are not one value but rather a membership to all possible classes with a specific membership level. To solve this problem, another user-defined variable is introduced "*alpha-class cut*" (α_c) [6]. This variable points to the value that has to be reached by class membership value of a record for this record to be considered a positive class instance. By using the variable "*alpha-class cut*" (α_c), it is possible to define positive (see Equation (2) left part) and negative (see Equation (2) right part) record sets:

$$P = \left\{ i \in T \middle| \mu_{concept}(i) \geq \alpha_c \right\}; N = \left\{ i \in T \middle| \mu_{concept}(i) \geq \alpha_c \right\} \qquad (2)$$

where P is the positive set of records for the corresponding class,

N – the negative set of records for the corresponding class,

i – the record from the training set,

T – the training data set,

$\mu_{concept}(i)$ – the membership value of *i-th* record to the corresponding class,

α_c – the alpha-class cut value,

concept – the corresponding class.

Before inspecting the real BEXA conditions in the context of FuzzyBEXA, it is important to note the fact that FuzzyBEXA algorithm does not contain the use of a specific membership function – in its place there is fuzzy data analysis. Since FuzzyBEXA algorithm uses fuzzy data, this algorithm does not differentiate between processing of categorical and continuous data [2].

B. FuzzyBEXA II

FuzzyBexa II – in this algorithm each class is not examined individually; instead, rules for all classes are generated. The highest level (Cover) does not split the training set into positive and negative sets; it transfers the whole training set and the set of concepts to the middle level procedure. The middle level procedure – find the best conjunction – finds both the conditional (antecedent, IF) and the resulting (consequent, THEN) part for each rule. Respectively, the lowest level procedure that generates specializations also processes the whole training set (or its part) instead of positive and negative instances of a split data set [3].

III. The Used Data Sets

This study uses 16 data sets that can be conditionally divided into three groups. Initially the classification algorithms are tested using popular UCI data sets (Iris data set, Auto MPG and Ionosphere data set [7]) to evaluate the results of these algorithms comparing them to other algorithms.

Then a series of experiments are carried out using real natural data available in the UCI repository to assess the accuracy of the algorithms using real medium-sized data sets (Nursery data set, Breast cancer Wisconsin, Parkinson, SPECT heart, Molecular biology (Splice-junction gene sequences), Yeast data set [7]).

The section of practical experiments is concluded with experiments that use real bioinformatics data sets (GSE3726 (Breast & colon cancer), GSE2535 (CML treatment), GSE2685 (Gastric cancer), GSE1577 (Lymphoma & Leukaemia), GSE2191 (AML prognosis), GSE89 (Bladder cancer) and GSE1987 (Lung cancer) [8]). The description of the data sets is given in Table I.

TABLE I
THE USED DATA SETS

Name	Number of samples	Number of attributes /genes	Number of classes
Iris data set (UCI)	150	4	3
Auto MPG data set (UCI)	398	8	2
Ionosphere data set (UCI)	351	34	2
Nursery data set (UCI)	12960	8	3
Breast Cancer Wisconsin (UCI)	699	10	2
Parkinson (UCI)	197	23	2
SPECT heart (UCI)	267	22	2
Molecular biology (Splice-junction gene sequences)(UCI)	3190	61	3
Yeast (UCI)	1484	8	10
GSE3726 (Breast & colon cancer)	52	22283	2
GSE2535 (CML treatment)	28	12625	2
GSE2685 (Gastric cancer)	30	4522	3
GSE1577 (Lymphoma & Leukaemia)	29	15434	3
GSE2191 (AML prognosis)	54	12625	2
GSE89 (Bladder cancer)	40	5724	3
GSE1987 (Lung cancer)	34	10541	3

IV. Experiments

All experiments using algorithms of BEXA family have been carried out in the application created using Java programming language (using Weka libraries [9]). All experiments include evaluation using cross-validation.

The experiment plan includes the data sets described in the previous section. To compare the results of BEXA family classification algorithms, experiments have been conducted using other popular algorithms and the same data sets using Weka [9] and Keel [10] software. It has been performed with the aim to ascertain the competitiveness of the classification algorithms and draw the necessary conclusions, as well as answer the raised question – is the use of BEXA family classification algorithms recommended for bioinformatics data classification and whether it has potential.

A. Bexa Classification Results

The obtained results show that the inductive classification algorithm BEXA achieves comparatively high classification accuracy results in categorical data ("Splice", "SPECT", "Breast Cancer", "Nursery") classification (see Fig. 2).

However, in the case of continuous data ("AML prognosis", "Brest colon cancer", "Gastric cancer", "Lymphoma & leukaemia"), its classification accuracy decreases (see Fig. 3). In the case of categorical data classification, BEXA algorithm shows better results in all data sets, comparing to PRISM algorithm. In some specific data sets, e.g. "SPECT", the classification accuracy of BEXA algorithm compared to PRISM algorithm is, on average, 12% higher (see Fig. 2). However, when compared to other algorithms (JRip and PART) BEXA algorithm achieves comparable classification accuracy results.

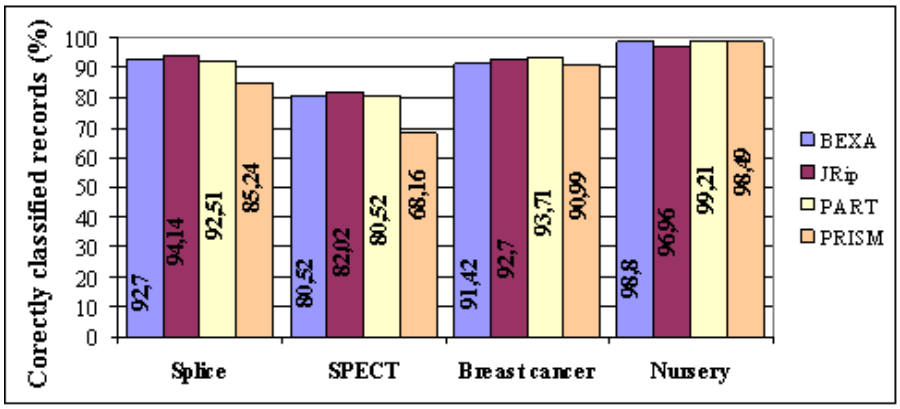

Fig. 2. Results of BEXA family algorithm classification – categorical data

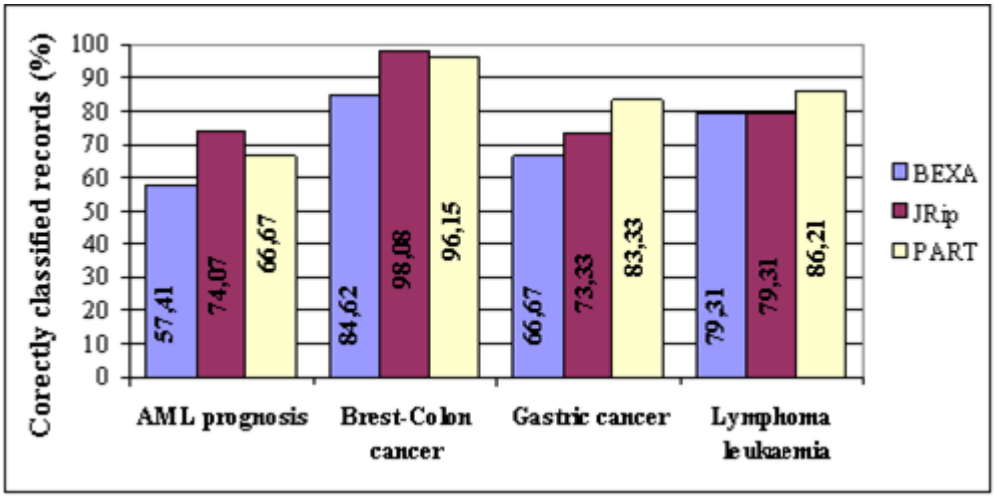

Fig. 3. Results of BEXA family algorithm classification – continuous data

B. Classification Results of FuzzyBEXA Algorithms

The classification accuracy of FuzzyBEXA algorithm is influenced by several parameters. This article studies the influence of the following parameters on the classification accuracy of FuzzyBEXA algorithm:

- the impact of alpha-cut variable on the results;
- the impact of the interval number in the triangle membership function on the results.

1. The Impact of Alpha-Cut Variable on Results

The influence of variable αI on the classification accuracy of algorithm FuzzyBEXA can be divided into three groups:

- The influence of variable αI is comparatively small and data are classified similarly disregarding its value. The impact of this variable αI can be observed in data sets ”Auto”, “Ionosphere” and “Parkinson”;
- A significant decline in classification results with αI values close to middle can be observed in data sets “Breast-Colon cancer”, “AML prognosis”, “Bladder cancer”, “Gastric cancer”, “Lymphoma & leukaemia”, and partly “CML treatment” and “Lung cancer” data sets can be added to this group of behaviour (see Fig. 4)
- Classification accuracy initially rises when the variable αI is increased, but reaching the highest point at the values between 0.3 and 0.4 classification accuracy decreases. This type of variable αI impact can be observed in data sets “Yeast” and “Iris” (see Fig. 4).

The classification accuracy of FuzzyBEXA algorithm has a tendency to initially slightly increase when the variable αT is increased (this trend can be seen in all data sets excluding “Ionosphere” data set). But when the variable αT reaches value 0.4, every data set shows individual behaviour (see Fig. 5).

2. The Impact of the Interval Number in the Triangle Membership Function on Results

The impact results of the membership function construction parameter “number of intervals” are presented in Fig 6. The results show that also the impact of interval number on correct classification has a very individual nature. Data sets that are almost not affected by the changes of this parameter are “Auto”, “Ionosphere”, “Iris” and “Yeast” (see Fig. 6). Data sets that show great response to changes of the parameter are, e.g. “Bladder cancer” and “Gastric cancer”.

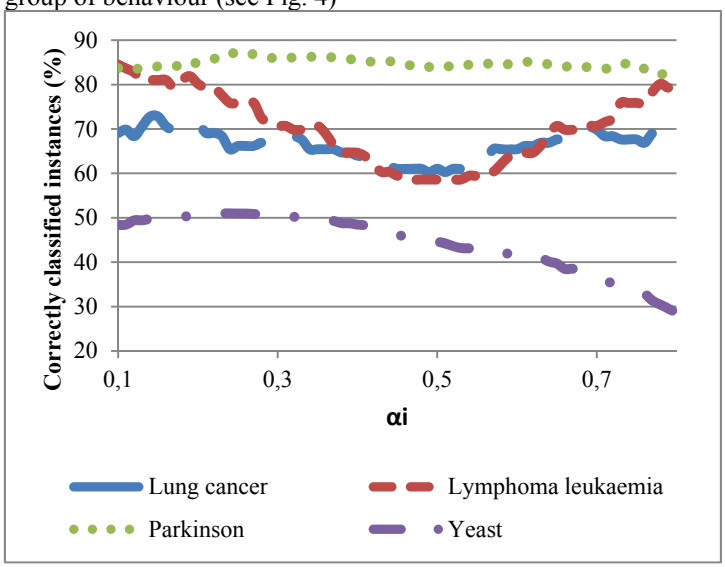

Fig. 4. Impact of alpha-cut variable (α_I) on classification accuracy

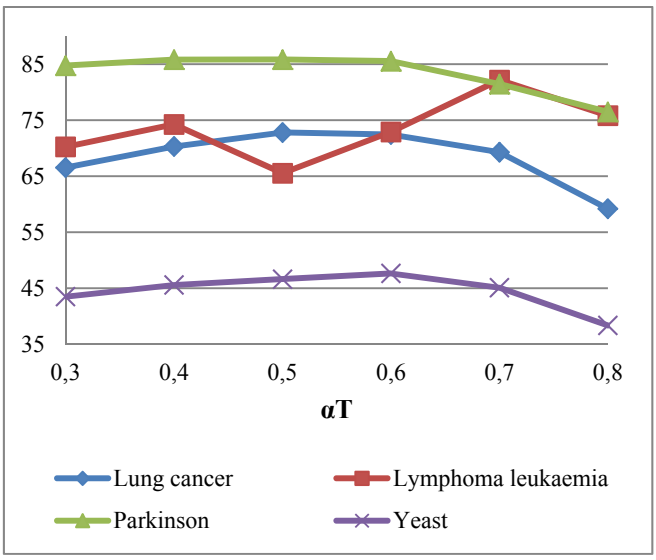

Fig. 5. Impact of alpha-cut variable (α_T) on classification accuracy

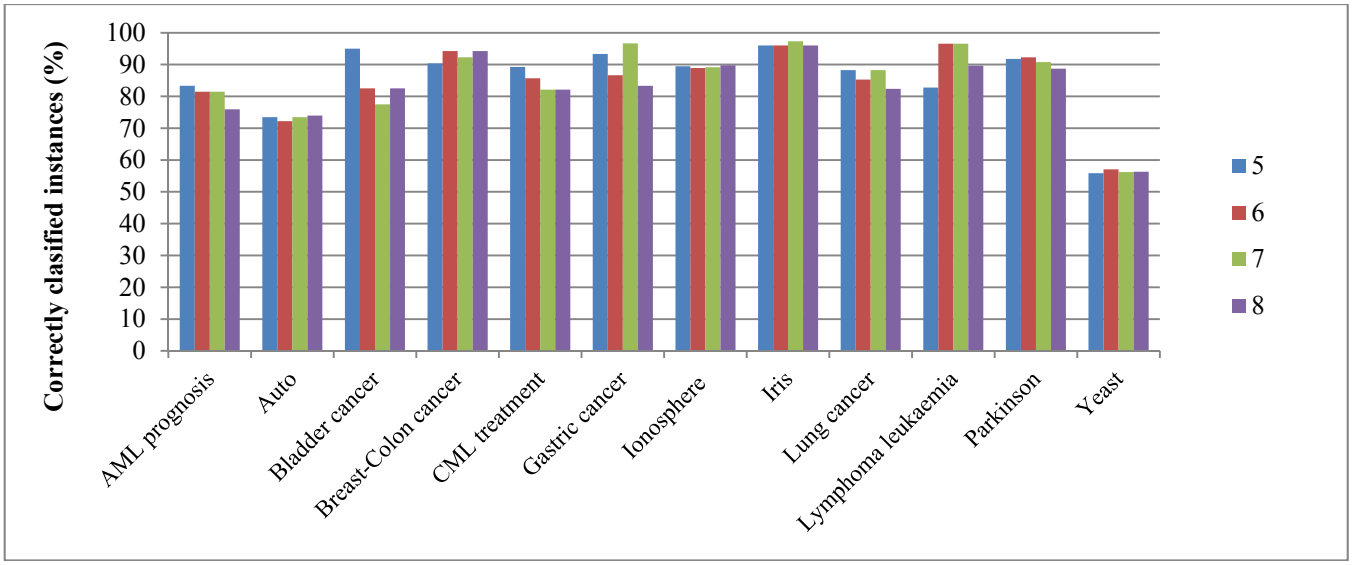

Fig. 6. The impact of the interval number in the triangle membership function on the results

TABLE II
CLASSIFICATION ACCURACY, %

Data set name	FuzzyBEXA	FURIA	FLR	SLAVE C	FuzzyBeXA II
Iris	97.33	94.67	91.33	95.33	56.00
Auto	73.98	81.38	66.84	69.88	72.45
Ionosphere	89.74	90.31	69.52	91.15	64.10
Yeast	57.08	58.09	28.10	49.73	48.05
Parkinson	92.31	87.69	82.56	84.53	88.21
AML prognosis	83.33	70.37	94.44	63.67	62.96
Breast-Colon cancer	94.23	96.15	88.46	87.00	82.69
Gastric cancer	96.67	80.00	83.33	66.67	86.67
Lymphoma & leukaemia	96.55	89.66	100.00	86.67	75.86
CML treatment	89.29	82.14	100.00	76.67	85.71
Bladder cancer	95.00	82.50	97.50	87.50	70.00
Lung cancer	88.24	88.24	94.12	80.00	97.06

3. Classification Accuracy

It is obvious that classification accuracy of all algorithms fluctuates in the range from 70% to 100%. However, it should be noted that "Yeast" data set shows considerably low classification accuracy results compared to other data sets (see Table II). It can be explained by the fact that this set is the only data set that holds 10 classes; furthermore, some classes have very poor support among instances. When FuzzyBexa parameters are applied to FuzzyBexa II algorithm, the obtained results are clearly worse (see Table II). It shows that both of these Bexa family fuzzy algorithms do not have the same best parameters; instead they should be experimentally determined individually for algorithm.

V. RECOMMENDATIONS

Based on the obtained results, recommendations about the application possibilities in real data classification can be given. In the case of FuzzyBEXA algorithm, the capacity of calculation resource workload is comparably high because processing of fuzzy data often asks for additional calculations during the execution of the algorithm, e.g., such necessity can be caused by belonging of a specific record to all linguistic values of an attribute. Inductive learning algorithm BEXA asks for less calculation resources and therefore it is a suitable data classification tool in the case if:

1. it is necessary to carry out classification of categorical data;
2. data can be categorized without losing important information.

However, in the real life often data categorization is not possible. An example is molecular biology data sets that describe gene expression levels. In this area data categorization can often be impossible because the influence of gene expression levels on different genetically determined indications is not fully researched and understood. Data categorization problem gives ground to broad application

potential for fuzzy data analysis, including the use of inductive learning algorithm FuzzyBEXA.

The study results have been obtained using several data sets that include molecular biology data sets (e.g., "Bladder cancer", "Breast-colon cancer", "Lymphoma & leukaemia", etc.). The classification accuracy results of these data sets show that FuzzyBEXA algorithm is potentially suitable for solving such tasks. However, it should be noted that to use the potential, there is a need for extensive experiments to find the optimal parameters, because at present there are no guidelines or strategies describing the choice of FuzzyBEXA parameters. The lack of such guidelines (that substantiates the need for the extensive experiments) is the most significant deficiency of FuzzyBEXA algorithm and the potential for further research.

Based on the results obtained in this study, all parameters and data sets can be compared; it is also possible to search for connections between FuzzyBEXA optimal parameters and data set specifics. This type of studies can help to form a unitary strategy for choosing FuzzyBEXA algorithm parameters. A unitary parameter choice strategy development asks for extensive experiments with this algorithm. The results obtained in this study can be viewed as the basis to reach this goal; but further research in this direction is necessary. To carry out further experiments one can use software developed in this study.

VI. Conclusions

The algorithms of BEXA family can be used in bioinformatics data classification, but the obtained results are not competitive when compared to other popular data mining algorithms. Additional experiments are necessary to improve classification results and assess the impact of various membership functions on the classification accuracy of BEXA family algorithms. Recommendations are given about the use of the most successful algorithm of BEXA family based on the used data set.

The obtained results allow drawing the following conclusions:

1. Inductive classification algorithm BEXA is a suitable tool for data classification if the classified data is categorical or can be categorized without losing significant information.
2. In general, inductive classification algorithm FuzzyBEXA shows statistically higher classification accuracy compared to algorithm SLAVE C. When FuzzyBEXA is compared to other algorithms used in the study, it is obvious that FuzzyBEXA algorithm achieves stable tendency to classify data with higher accuracy.
3. Overall, it can be concluded that inductive classification algorithm FuzzyBEXA has high application potential in the case of using optimal parameters.
4. Inductive classification algorithm FuzzyBEXA shows tendency to classify data worse, when alpha-cut (αI) values are near the middle of the range (around 0.5).
5. Inductive classification algorithm FuzzyBEXA classification accuracy increases, when alpha-cut (αT) value is increased to 0.4.
6. In the case of inductive classification algorithm FuzzyBEXA, the number of membership function intervals does not influence classification accuracy.

Acknowledgments

Thanks to Dr.habil.sc.comp. Professor Arkady Borisov for help and support.

References

[1] H. Theron, I. Cloete, BEXA: A Covering Algorithm for Learning Propositional Concept Descriptions, in Machine Learning, Vol. 24, Boston: Kluwer Academic Publishers, 1996, pp.5-40.

[2] J. van Zyl, I.Cloete, FuzzConRi – A Fuzzy Conjunctive Rule Inducer, in Proc. Workshop on Advances in Inductive Rule Learning, ECML, 2004, pp.194-203.

[3] J. van Zyl, I.Cloete, Simultaneous Concept Learning of Fuzzy Rules, in Proc. Workshop on Advances in Inductive Rule Learning, CCML, 2004, pp.194-203.

[4] P. Clark. The CN2 Induction Algorithm / Clark P. and Niblett T. // Machine Learning. Vol. 3, 1989, pp. 261-283.

[5] J. Hong. AQ15: Incremental Learning of Attribute-Based Descriptions from Examples the Method and User Guide. Report of the Intelligent Systems Group, UIUCDCS-F-86-949 Department of Computer Science, University of Illinois at Urbana-Champaign, Urbana, IL, 1986.

[6] J. Zyl. Fuzzy set covering as a new paradigm for the induction of fuzzy classification rules. – Mannheim: PhD thesis, 2007. p 263.

[7] A. Frank, A. Asuncion, UCI Machine Learning Repository Irvine, CA: University of California, School of Information and Computer Science. 2010. [Online] Available: http://archive.ics.uci.edu/ml]. [Accessed: June 3, 2012]

[8] M. Gasparoviča M., L. Aleksejeva. Feature Selection for Bioinformatics Data Sets – Is It Recommended? // Proceedings of the 5th International Conference on Applied Information and Communication Technologies (AICT2012), Latvia, Jelgava, 26.-27. April, 2012. - pp 325-335.

[9] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, I.H.Witten.: The WEKA Data Mining Software: An Update. SIGKDD Explorations. 11:1, 2009, pp. 10-18.

[10] J. Alcalá-Fdez,, A. Fernandez, J. Luengo, J. Derrac, S. García, L. Sánchez,, F. Herrera: KEEL Data-Mining Software Tool: Data Set Repository, Integration of Algorithms and Experimental Analysis Framework. Journal of Multiple-Valued Logic and Soft Computing. 17:2-3, 2011, pp.55-287.

Madara Gasparoviča received her diploma of Mg. sc. ing. in Information Technology from Riga Technical University in 2010. Now she is a doctoral student at the study programme "Information Technology", Riga Technical University.

Since 2008 she has worked as a Senior Laboratory Assistant at Riga Technical University, and since 2010 she has been working as a Researcher at the Department of Modelling and Simulation, the Institute of Information Technology. Previous publications: Gasparovica M., Novoselova N., Aleksejeva L., *Using Fuzzy Logic to Solve Bioinformatics Tasks*, Proceedings of Riga Technical University. Issue 5, Computer Science. Information Technology and Management Science, Vol.44, 2010, pp.99-105. Gasparoviča M., Aleksejeva L. *Using Fuzzy Unordered Rule Induction Algorithm for Cancer Data Classification*, Proceedings of the 17th International Conference on Soft Computing, MENDEL 2011, Czech Republic, Brno, June 15-17, 2011, pp. 141-147.

Her interests include decision support systems, data mining tasks and modular rules. She is a member of IEEE.

Address: 1 Kalku Street, LV-1658, Riga, Latvia. E-mail: madara.gasparovica@rtu.lv.

Ludmila Aleksejeva received her Dr. sc. ing. degree from Riga Technical University in 1998. She is an Associate Professor at the Department of Modelling and Simulation, Riga Technical University. Her research interests include decision making techniques and decision support system design principles, as well as data mining methods and tasks, and especially collaboration and cooperation of the mentioned techniques.

Most important previous publications: Gasparoviča M., Novoselova N., Aleksejeva L., *Using Fuzzy Logic to Solve Bioinformatics Tasks*, Proceedings of Riga Technical University. Issue 5, Computer Science. Information Technology and Management Science, Vol.44, 2010, pp.99-105. Gasparoviča M., Aleksejeva L., Tuleiko I. *Finding Membership Functions for Bioinformatics Data* // Proceedings of the 17th International Conference on Soft Computing, MENDEL 2011, Czech, Brno, June 15-17, 2011.pp. 133-140.

Address: 1 Kalku Street, LV-1658, Riga, Latvia. E-mail: ludmila.aleksejeva_1@rtu.lv.

Valdis Gersons received his Bachelor Degree in Information Technology from Riga Technical University in 2012. He elaborated his Bachelor Thesis on the inductive methods in bioinformatics data classification.

Address: 1 Kalku Street, LV-1658, Riga, Latvia. E-mail: valdis.gersons@rtu.lv.

Madara Gasparoviča, Ludmila Aleksejeva, Valdis Gersons. Bexa saimes algoritmu pielietošana bioinformātikas datu klasifikācijā

Šajā rakstā pētītas Bexa saimes algoritmu iespējas reālu bioinformātikas datu klasifikācijā. Bexa saime sastāv no trim algoritmiem: Bexa – kas darbojas ar stingriem datiem, kā arī FuzzyBexa un FuzzyBexa II, kas darbojas ar izplūdušiem datiem. FuzzyBexa no FuzzyBexaII atšķiras ar to, ka pēdējā katra klase netiek apskatīta individuāli, bet gan tiek ģenerēti likumi visām klasēm. Bexa saimes algoritmi nosacīti sastāv no trim daļām –pārklājuma procedūras, labākā likuma meklēšanas, izmantojot novērtējuma funkciju, kā arī specializāciju veidošanas. Praktiskie eksperimenti tika veikti ar sešpadsmit reālām datu kopām, kuras nosacīti var iedalīt trīs daļās: literatūrā bieži izmantotās datu kopas (Iris data set, Auto MPG and Ionosphere Data Set), UCI datu repozitorija reālas bioinformātikas datu kopas (Nursery Data Set, Breast cancer Wisconsin, Parkinsons, SPECT heart, Molecular biology (Splice-junction gene sequences), Yeast data set) un reālas bioinformātikas datu kopas, kam ir liels atribūtu un mazs ierakstu skaits (GSE3726 (Breast & colon cancer), GSE2535 (CML treatment), GSE2685 (Gastric cancer), GSE1577 (Lymphoma & Leukaemia), GSE2191 (AML prognosis), GSE89 (Bladder cancer) and GSE1987 (Lung cancer)). Lai salīdzinātu Bexa saimes algoritmu klasifikācijas rezultātus, tika veikti papildus eksperimenti ar visām izmantotajām datu kopām ar citiem algoritmiem: Bexa klasifikācijas rezultāts kategoriskiem datiem salīdzināts ar JRIP, Part un PRISMA algoritmiem, kā arī ar skaitliskiem datiem ar Jrip un Part. FuzzyBexa un FuzzyBexaII klasifikācijas rezultāti salīdzināti ar FURIA, FLR un Slave C algoritmiem. Pēc klasifikācijas rezultātiem izdarīti secinājumi par atsevišķu kritēriju ietekmi uz iegūto klasifikācijas rezultātu. Pēc rezultātiem redzams, ka šīs saimes algoritmu izmantošana bioinformātikā ir perspektīva un nepieciešami tālāki pētījumi par iespējām uzlabot algoritmu vājās puses, lai paaugstinātu to klasifikācijas precizitāti un iegūto likumu kvalitāti.

Мадара Гаспаровича, Людмила Алексеева, Валдис Герсонс. Применение алгоритмов семейства Веха в классификации данных биоинформатики

В данной статье исследуются возможности алгоритмов семейства Веха для классификации реальных данных биоинформатики. Семейство Веха состоит из трёх алгоритмов: Веха – который работает с чёткими данными, а также FuzzyBexa и FuzzyBexa II, которые работают с нечёткими данными. FuzzyBexa отличается от FuzzyBexa II тем, что в последнем каждый класс не рассматривается индивидуально, но генерируются законы для всех классов. Алгоритмы семейства Веха условно состоят из трёх частей: процедуры перекрытия, поиска лучшего закона, используя оценочную функцию, а также образования специализаций. Практические эксперименты проводились на шестнадцати реальных множествах данных, которые условно можно разделить на три части: часто используемые в литературе множества данных (Iris data set, Auto MPG и Ionosphere Data Set), реальные множества данных биоинформатики из репозитория данных UCI (Nursery Data Set, Breast cancer Wisconsin, Parkinsons, SPECT heart, Molecular biology (Splice-junction gene sequences), Yeast data set) и реальные множества данных биоинформатики, у которых большое количество атрибутов и маленькое количество записей (GSE3726 (Breast & colon cancer), GSE2535 (CML treatment), GSE2685 (Gastric cancer), GSE1577 (Lymphoma & Leukaemia), GSE2191 (AML prognosis), GSE89 (Bladder cancer) и GSE1987 (Lung cancer)). Чтобы сравнить результаты классификации алгоритмов семейства Веха, были проведены дополнительные эксперименты на всех использованных множествах данных с другими алгоритмами: результат классификации Веха для категорийных данных сравнён с алгоритмами JRIP, Part и PRISMA, а также для численных данных – с Jrip и Part. FuzzyBexa и FuzzyBexaII сравнены с алгоритмами FURIA, FLR и Slave C. По результатам классификации были сделаны выводы о влиянии отдельных критериев на полученный результат классификации. Исходя из полученных результатов классификации видно, что использование данного семейства алгоритмов в биоинформатике является перспективным, и необходимы дальнейшие исследования в контексте возможностей улучшить слабые стороны этих алгоритмов с целью повысить их точность классификации и качество полученных законов.