Silvia Trif, Adrian Visoiu / International Journal of Engineering Research and Applications (IJERA) ISSN: 2248-9622 www.ijera.com
Vol. 3, Issue 3, May-Jun 2013, pp.1085-1089

Gene Expression Programming Based Dataset Decoration for Improved Churn Prediction

Silvia Trif*, Adrian Visoiu**
*(Department of Doctoral Studies of Academy of Economic Studies, Bucharest, Romania)
** (Department of Economic Informatics and Cybernetics, Academy of Economic Studies, Bucharest, Romania)

ABSTRACT

Mobile network operators rely on business intelligence tools to derive valuable information regarding their subscribers. A key objective is to reduce churn rate among subscribers. The mobile operator needs to know in advance which subscribers are at risk of becoming churners. This problem is solved with classification algorithms having as input data derived from the large volumes of usage details recorded. For certain categories of subscribers, available data is limited to call details records. Using this primary data, a dataset is created to be conveniently used by a classification algorithm. Classification quality using this initial dataset is improved by a proposed method for dataset decoration. Additional attributes are derived from the initial dataset through generation, based on gene expression programming. Classification results obtained using the decorated dataset show that the derived attributes are relevant for the studied problem.

Keywords - *business intelligence, churn prediction, classification, data mining, gene expression programming*

I. INTRODUCTION

Nowadays, due to competition, one major challenge for service suppliers is to keep their existing position on market, if not gaining more market share. For the supplier, this creates the necessity to find ways to better address customer needs. The supplier has to understand the customer in order to model its behavior. Since quantitative information is objective and may be subject to further processing, any available customer data is collected for further analysis. This is also supported by the existence of information systems and data in electronic format. Datasets are created and fed to business intelligence specialized tools in order to obtain relevant information to support the decision making process. This processing flow enables a range of applications such as: churn prediction, suggestion engine, customer segmentation. Business Intelligence tools are not standalone, but integrated with the information system of the organization, such as CRM, as shown in [1]. Some applications of the intelligent engine consist in using Business Intelligence functionality directly from another system, without human intervention, like the platform for secure USSD service presented in [2]. The interactions between Business Intelligence tools and other systems may become complex. A reference complex churn prediction system is presented in [3].

II. PROBLEM FORMULATION

One of the problems service providers face is related to preventing churning. Churning occurs when a particular subscriber stops using the services provided by the supplier. The reasons for churning are various and the service provider may not have sufficient data to infer them. However, statistical analysis is able to discover patterns associated to churning behavior. When performing the analysis, past data is analyzed and patterns or rules are derived to identify churners. Statistical methods are used to estimate the churn risk for each subscriber as presented in [4]. When using in production, for existing subscribers, if they match churning patterns or rules, they are considered as potential churners and the service provider is able in advance to take preventive measures to stimulate the subscriber. Preventive measures include bonuses, discounts, gifts and other incentives that are likely to convince the subscriber not to churn and continue the relationship with the service provider [5].

In the following, we consider the case of a telecom operator as a service provider. Telecom operators have: a large number of subscribers of two main types – prepaid and postpaid - and a large number of records regarding service usage, stored in files or databases. An important step in the creation of the dataset is gathering and transforming primary data to the format needed by the algorithm used in analysis. Standard approaches to data mining include the usage of data warehouses as presented in [6]. Churn analysis results quality is influenced by certain factors: available data in terms of volume and details or attributes and algorithms used.

Regarding classification algorithms, the reference algorithm used in analysis is Naïve Bayes.

Several other algorithms fit to solve this problem are decision trees and logistic regression as presented in [7].

Regarding available data at telecom operators, for both prepaid subscribers there are call detail records showing when and how the services were consumed. There are also differences between prepaid and postpaid subscribers regarding supplemental information regarding the subscriber. Postpaid subscribers consume their service based on a sign contract and therefore the operator has access to identification data such as name, gender, age and similar. This kind of information is important in churn analysis as it represents additional attributes helping the algorithm to give better results. On the other hand, prepaid subscribers are anonymous for the service provider. The only identification is the phone number – MSISDN - they have on the prepaid SIM card. Therefore, churn prediction is a lot more difficult for prepaid subscribers for which only service consumption is known. Based on the service consumption, for each MSISDN, service usage is summed up for a number of equidistant and consecutive periods or time in order to build the initial dataset. A record has the form of MSISDNi, USGi1, USGi2, …, USGij, …, USGik, STATE, where:

- MSISDNi is the identifier of the prepaid subscriber in the system
- USGij is the amount of service consumed by subscriber i during period j
- STATE is a Boolean value showing if the subscriber was a churner, based on the amount consumed by the subscriber i during the k+1 period, if USGik+1>0 then the subscriber is "active", if USGik+1=0 then the subscriber is "inactive"

The set of records defined above constitutes the initial dataset. In this context, the churn prediction analysis aims to estimate the value of STATE variable depending on USG_1, USG_2, .., USG_k. This is treated as a classification problem and classification algorithms are applied. However, it is observed that simply employing the initial dataset for analysis leads to poor classification results as shown in [8]. Dataset decoration is needed to add relevant attributes that would improve the quality of the results. In case of poor information available, decoration is done using derived attributes that are obtained by transforming the initial attributes such way to:

- Eliminate unnecessary correlations between initial attributes
- Improve the dependency between the derived attributes and the studied variable
- Create meaningful derived attributes, easy to compute and to interpret
- Gain more information from the derived attributes than from initial attributes

In [8] a transformation for creating derived attributes based on predefined basic statistical indicators is presented. In [9] more attributes are derived by analyzing the call graph. In our research we considered only the above defined consumption indicators.

In the following we investigate a free transformation of the above initial dataset based on expression generation using a dedicated evolutionary algorithm, the gene expression programming.

III. GENE EXPRESSION PROGRAMMING BASED DATASET DECORATION

In the following, the simple transformations made on the initial attributes are extended to a more general approach. Using simple operators like addition, subtraction, multiplication, division as well as operands taken from the attribute set, then a large number of expressions may be obtained. As the number of expression forming combinations is very large, and generating all the possible combinations is time consuming, a solution taken into account to obtain a valid result in a short time is to use gene expression programming as a mean of generating derived attributes based on the set of initial attributes and a set of operators to be applied.

Gene expression programming is an evolutionary algorithm based on the principles described in [10]. A population of chromosomes evolves through applying genetic operators iteratively. Genetic operators include: elitism – survival of the best chromosomes from a generation, crossover – exchange of genetic material between two chromosomes, mutation – random change inside existing chromosomes. Chromosomes encode, symbolically, individuals representing solutions of a problem to be solved. The design and preparation of the genetic algorithm in general, and of the gene expression programming, in particular, requires that specific concepts from the evolutionary algorithm need to be put in correspondence with elements of the problem to be solved.

Structurally, chromosomes specific to gene expression programming are made up of one or more genes. Genes are made up of symbols. Chromosomes encode expressions made up of operands and operators. Due to flexibility needs, chromosomes allow the encoded expression to be obtained by combining several sub-expressions encoded by genes, such way each gene encodes a sub-expression.

A symbol encodes an operand or an operator. Since the expressions encoded by genes are intended to be evaluated, there are specific genetic expression programming rules to correctly form a gene. A gene is made up of a head having symbols encoding both operators and operands, and a tail which is made up only of operands. The length of the head and the length tail take into account the maximum number of operands needed by the operators taken into account, such way the syntax tree associated to the expression to be well formed, as presented in [11].

In our research, we aim to obtain a dataset having attributes derived from initial attributes, under the form of analytical expressions made up of operands and operators. Taking into account the structure of the gene expression programming chromosome, we propose the following mapping between gene expression programming concepts and our problem. Each chromosome encodes the list of derived attributes. Each gene of the chromosome encodes the expression of the derived attribute. Each symbol encodes an operator from a predefined set or an operand taken from the set of initial attributes. In the following, we consider the initial dataset containing only usage attributes: USG_1, …, USG_4.. Table 1 presents a gene encoding for the expression, while Fig. 1 shows the actual syntax tree inferred from the encoding, in a top-down, left-to-right fashion.

E= (USG3-USG2)/USG1

Gene:

Table 1 A gene encoding

head		tail		
/	-	USG1	USG3	USG2

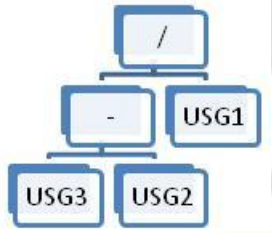

Fig. 1 Syntax tree for the gene

For example, a two-gene chromosome, with genes having a head length of 2 positions and a tail length of 3 positions is presented in Fig. 2:

Chromosome:

Gene position	0					1					
Genes	G1					G2					
	position	0	1	2	3	4	position	0	1	2	3	4
	symbol	S1	S2	S3	S4	S5	symbol	S6	S7	S9	S9	S10

Fig. 2 Two gene chromosome

The preparation of the gene expression programming approach requires:

1) Setting up the list of operators, OT, and the list of operands, OD. For example, the list of operators, OT is OD={+,-,*,/} and the list of operands, OD is OD={USG1,USG2, USG3, USG4}

Crossover operation that occurs between two randomly chosen chromosomes enables the exchange of genetic material in order to obtain new individuals with new characteristics. Crossover occurs at several positions inside the chromosome. The crossover between two single gene chromosomes at a certain position is presented in figure below:

Chromosome 1, encoding the expression (USG2-USG3)+USG1 is presented in Fig. 3:

Position		0	1	2	3	4
Symbol		+	-	USG1	USG2	USG3

Fig. 3 Chromosome1before crossover

Chromosome 2, encoding the expression (USG3+USG4)-USG2 is presented in Fig.4:

Position		0	1	2	3	4
Symbol		-	+	USG2	USG3	USG4

Fig. 4 Chromosome2 before crossover

After crossover at position 3, chromosome 1', encoding the expression: (USG3-USG4)+USG1 is presented in Fig. 5:

Position		0	1	2	3	4
Symbol		+	-	USG1	USG3	USG4

Fig. 5 Chromosome1' after the crossover

Chromosome 2', encoding the following expression (USG2+USG3)-USG2, is presented in Fig. 6:

Position	0	1	2	3	4
Symbol	-	+	USG2	USG2	USG3

Fig. 6 Chromosome 2' after the crossover

The syntax trees associated to the initials chromosomes before and after crossover are presented in Fig. 7 and Fig. 8, respectively:

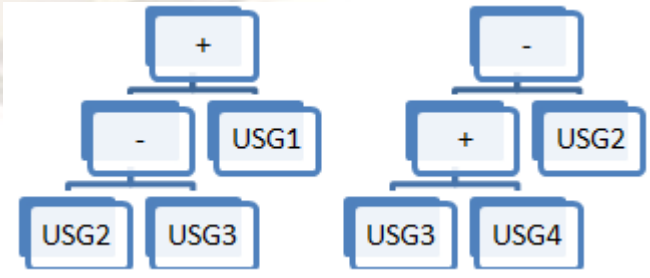

Fig. 7 Expression trees before crossover, corresponding to chromosomes 1 and 2, respectively.

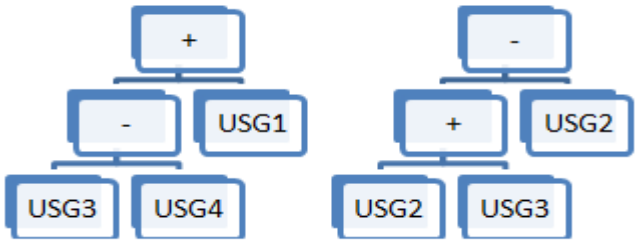

Fig. 8 Expression trees after crossover, corresponding to the chromosomes 1' and 2'. respectively.

This kind of crossover may happen in several positions, affecting gene contents in both single-gene and multiple-gene chromosomes. This enables the evolution of expressions of the derived attributes through the implied changes.

A special case of crossover is when such exchange of genetic material occurs only at positions pointing to the beginning of genes, inside a multiple-gene chromosome. In this case, entire attributes encoded by genes will be exchanged, as shown in Fig. 9 and Fig. 10, respectively.

Chromosome 1:

Position	0	1	2	3	4
Gene	G11	G12	G13	G14	G15

Chromosome 2:

Position	0	1	2	3	4
Gene	G21	G22	G23	G24	G25

Fig. 9 Multi-gene chromosomes before crossover at gene position

After crossover at position 2 the new chromosomes are the ones presented in Fig. 10:

Chromosome 1':

Position	0	1	2	3	4
Gene	G11	G12	G23	G24	G25

Chromosome 2':

Position	0	1	2	3	4
Gene	G21	G22	G13	G14	G15

Fig. 10 Multi-gene chromosomes after crossover at gene position

Entire gene crossover enables the evolution of derived attributes lists, through the change they imply. Both of these two kinds of the same genetic operator may be applied to the chromosome population independently.

Mutation is a sudden change in the content of a gene, randomly changing a symbol. For example, Fig. 11 shows a single gene chromosome before the mutation of an operator that changes the encoded expression.

Chromosome 1: (USG2-USG3)+USG1

Position		0	1	2	3	4
Symbol		+	-	USG1	USG2	USG3

Fig. 11 Chromosome 1 before mutation

After mutation at position 1, changing minus sign with plus, the chromosome becomes as presented in Fig.12:

Chromosome 1': (USG2+USG3)+USG1

Position		0	1	2	3	4
Symbol		+	+	USG1	USG2	USG3

Fig. 12 Chromosome 1 after mutation

Using the proposed gene expression programming approach, an initial population of multi-gene chromosomes has been evolved in order to obtain a list of derived attributes corresponding to the best chromosome after a threshold number of iterations. The performance criteria was the quality of the classification with respect to each attribute encoded by a chromosome.

The list of derived attributes obtained using gene expression programming is shown in table 2.

Table 2 The list of attributes derived using gene expression programming

Gene	Encoded expression
1	USG1+USG2+USG3+USG4
2	(USG2-USG1)/USG1
3	USG3+USG4
4	(USG3+USG4)/USG1
5	(USG2+USG3)/USG1
6	USG4-USG3
7	(USG4-USG3)/(USG1+USG2)
8	(USG3-USG1)/(USG1+USG2)

Classifying the subscribers using the whole derived dataset, the percent of correctly classified instances is 58.56% while the incorrectly classified instances are 41.13% of the set, as obtained from table 3.

Table 3 Correctly and incorrectly classified subscribers using the list of derived attributes

	Classified as active	Classified as inactive
Actual active	5595	7315
Actual inactive	972	6118

This is an improvement over the results obtained using the initial usage attributes. Further improvement is made by generating combinations of attributes and asses the quality of the classification for each individual combination. The combination of attributes having the best classification quality is chosen. Table 4 shows the derived attributes chosen as best combination in our case study.

Table 4 The combination of derived attributes that is best for the classification

No	Derived attribute
1	(USG3+USG4)/USG1
2	(USG2+USG3)/USG1
3	USG4-USG3

Using the dataset defined by the attributes from table 10, the correctly classified instances are 71.50% of the total, while incorrectly classified instances are only 28.50%. The classification details are presented in table 5.

Table 5 Correctly and incorrectly classified subscribers using the best derived attributes

	Classified as active	Classified as inactive
Actual active	9639	3271
Actual inactive	2429	4661

It is also important that the chosen derived attributes have meaningful expressions with respect to the domain:

- (USG3+USG4)/USG1 shows the amount of usage for the last two weeks over the usage during initial week of the study; this is a relative indicator showing how many times the usage was higher with respect to the baseline
- (USG2+USG3)/USG1 is also a relative indicator showing how many times the usage was higher in a past period with respect to the baseline
- USG4-USG3 shows the absolute difference in consumption during consecutive periods.

Therefore, the generated attributes have statistical meaning, being easy to understand and interpret. They are also selected from a large set of generated expressions otherwise hard to assess without the convergence of evolutionary approach.

As seen, gene expression programming is a powerful method to use in order to obtain derived attributes from the original dataset.

IV. CONCLUSION

Dataset decoration is a useful technique for improving the quality of classification when poor data is available. We have proposed a dataset decoration method based on gene expression programming. For churn prediction analysis attributes derived using the proposed method are relevant and their use for classification brings significant improvement in the results of the classification.

ACKNOWLEDGEMENTS

This work was co-financed from the European Social Fund through Sectoral Operational Programme Human Resources Development 2007-2013, project number POSDRU/107/1.5/S/77213 „Ph.D. for a career in interdisciplinary economic research at the European standards".

REFERENCES

[1] A. Tudor, A. Bara, I. Botha, Data Mining Algorithms and Techniques Research in CRM Systems, *Proc. 13th WSEAS International Conference on Mathematical Methods, Computational and Intelligent* Systems, Iasi, Romania, 1-3 July 2011, 265-270.

[2] S. Trif, A. Vişoiu, USSD based one time password service , *Proc. 5th International Conference on Security for Information Technology and Communications* 2012, Bucharest, 2012, 141 -149.

[3] S. Yuan-Hung, David C. Yen, Hsiu-Yu Wang, Applying Data Mining to Telecom Churn Management, *Expert Systems with Applications*, 31(3), 2006, 512-524.

[4] R. J. Jadhav, Churn Prediction in Telecommunication Using Data Mining Technology, *International Journal of Advanced Computer Science and Applications*, 2(2), 2011, 87-110.

[5] J. Sathyan, M. Sadasivan, Next Generation Mobile Care Solution , *Proc 10th WSEAS International Conference on Telecommunications and Informatics*, Lanzarote, Canary Islands, Spain, May 27-29 2011, 265-270

[6] D. Camilovic, D. Becejski-Vujaklija, N. Gospic, A Call Detail Records Data Mart: Data Modeling and OLAP Analysis, *Computer Science and Information Systems*, 6(2), 2009, 17-19.

[7] Li-Shang Yangi , C. Chiu, Knowledge Discovery on Customer Churn Prediction", *Proc. 10th WSEAS Interbational Conference on Applied Mathematics*, Dallas, Texas, USA, November 1-3, 2006, 523-528.

[8] S. Trif, A. Visoiu, Improving Churn Prediction in Telecom through Dataset Decoration , *Proceding 7th WSEAS International Conference on Computer Engineering and Applications (CEA '13)*,9-11 Ianuarie 2013 Milano, Recent Researches in Information Science and Applications, Recent Advances in Computer Engineering Series, Vol. 9 , WSEAS Press, 223 – 228

[9] K. Dasgupta, R. Singh, et al., Social Ties and their Relevance to Churn in Mobile Telecom Networks, *Proc. 11th international conference on Extending database technology: Advances in database technology*, 2008, 668-677.

[10] B. C. Ferreira, *Gene Expression Programming: Mathematical Modeling by an Artificial Intelligence 2nd Edition* (Springer Publishing, May 2006).

A. Visoiu, Deriving Trading Rules Using Gene Expression Programming, *Informatica Economica*, 15(1), 2011, 22-30, ISSN 1453-1305