Fourth International Conference
Modelling and Development of Intelligent Systems
October 28 - November 1, 2015
"Lucian Blaga" University
Sibiu - Romania

Top - Down clustering used in analysis of the Romanian Teachers' Training Needs on Information and Communications Technology

Daniel Hunyadi, Daniel Mara

Abstract

This article presents some important aspects regarding the analysis on Romanian Teachers' Training Needs on Information and Communications Technology (ICT), using top-down clustering. The scope of this analysis is to create clusters regarding to the teacher's training needs on ICT. This analysis was made inside a project which generates significant results in the life-long learning of the teachers from all levels of the Romanian education system. The general objective of the project aims to increase the level of the teaching staff information, competences and abilities concerning the Information and Communications Technology. It also aims to improve the e-learning interactive methods and the activity with the disabled in order to use them also within the didactic activity, to improve the results of the educational process as well as to increase the disabled access to education.

1 Introduction

Education and training in Information and Communication Technology (ICT) are crucial in three of the seven flagship initiatives developed under the Europe 2020. "Digital Agenda for Europe" shows that the current information era requires increasing both digital literacy development and students' inclusion and it emphasizes the importance of access to education through ICT.
The strategy "Romania of Education, Romania of Research" establishes the need for directing education towards eight key skills "required for personal development and knowledge economy", including the acquisition and development of "digital literacy", and also insists on the need to digitize the curriculum.

The strategy "Education and Research for Knowledge Society" sets not only infrastructure-related commitments ("providing all schools with computers connected to the Internet and educational software able to raise the teaching and learning quality"), but also students' acquisition of digital skills and the need for "teachers' lifelong training in blended learning mode", as well.

The existence of such needs is supported by data and studies. Eurostat shows that the ability to use a computer and the Internet are still very low in Romania. In 2012 (the latest data available), only 13% of the population was able to carry out 3 or 4 tasks on the Internet (out of 6 tested tasks) and 17% carried out 3-4 computer-specific tasks (out of 6), Romania being the last of the 31 countries surveyed (compared to the EU 27 average – of 25% and 27% respectively).

The study "Survey of Schools: ICT in Education" ordered by the European Commission shows that Romanian teachers' computer literacy (especially in secondary education) is below the European average.

Regarding the teacher's role in the web age, one may say without any doubt that this is in a continuous transformation process. By the nature of their job, the teacher should mediate and facilitate student's knowledge and training, but without neglecting aspects such as: students' developing critical thinking, increasing communication and networking with peers, and working collaboratively.

If we were to overlap all these issues with web features, we can see that they are compatible, to a large extent. In other words, through technology, applications and web services the instruction process may benefit from the basic elements characterising education, namely: users' interaction and communication, sharing and cooperation established between multiple users, and the information and training processes, as well.

Therefore, to maximize these benefits during the educational process the teacher must know, get acquainted to and be able to exploit web applications, services and technologies. Whether wikis, blogs, podcasts, social networking, bookmarking tools, labelling or social annotation sites, information syndicating processes, specialized search-engines, widgets/ gadgets and so on, they can be extremely useful to teachers in training students.

The major problem which prevents generalization of using these tools by teachers in the educational process is the lack of adequate time for appropriate training. With few exceptions, which are not characteristic for Romanian education, teachers (regardless of the subject they teach) do not receive help from qualified and authorized persons in preparing and using these tools. Therefore, the workload of a teacher should be huge, detrimental to his/her other tasks. Equally true is the fact that, the benefits and satisfaction are worth, so that teachers should not miss the opportunities arisen in lifelong learning.

A teacher in the current era must understand that the student is in the centre of the instruction process, and he/she, as a teacher, should facilitate all these and create an educational environment for the student to be able to access applications, services and current technologies appropriate for the educational process. The teacher's role is also to guide pupils and students, to encourage them to get engaged in conversations, whether virtual or real, both with peers and with teachers, the teacher-student role being thus often reversed. The teacher must therefore understand that he/she shall gain a lot from his/her students' learning experiences, and that they shall go together on this way of reducing physical boundaries of a classroom setting.

More than ever, teachers should guide the students in understanding and critically analysing the information content accessed, in terms of quality and accuracy, especially since the amount of information often grows uncontrollably.

Young people nowadays are using much of their free time to carry out online activities, such as: creating digital content (by posting on blogs, participating in social networks, etc.), communicating with friends/ colleagues/ family, looking for information or creating educational materials. Despite these digital skills and ICT-mediated communication or virtual activities for recreation and leisure, youth are quite poorly prepared when they need to use these skills for academic purposes. When they need to create educational material, gaps on selecting and synthesising the materials they have found are visible from the very first moment. Similarly, critical observation from a comparative perspective is another element that would ultimately lead to less satisfactory results.

2 Literature review

2.1 Clustering and k-means Algorithm

According to [2], clustering is the unsupervised classification of patterns (observations, data items, or feature vectors) into groups (clusters). Vaishali opined that clustering algorithms generate clusters having similarity between data objects based on some characteristics [5].

Clustering is extensively used in many areas such as pattern recognition, computer science, medical, machine learning. Jean Yan states that "formally cluster structure is represented as a set of subset C=C1,……..Ck of S, such that S=$\cup_{i=1}^{k}$ Ci and $C_i \cap$Cj=∅ for $i \neq j$. Consequently, instances in S belong to exactly one and only one subset". Clustering algorithms have been classified into hierarchical and partitional clustering algorithms. Hierarchical clustering algorithms create clusters based on some hierarchies. It is based on the idea of objects being more related to nearby objects farther away [2]. It can be top-down or bottom-up hierarchical clustering. The top-down approach is referred to as divisive while the bottom-up approach is known as agglomerative. The partitional clustering algorithms create various partitions and then evaluate them by some criterion. k-Means algorithm is one of most popular partitional clustering algorithm. It is a centroid-based algorithm in which each data point is placed in exactly one of the K non-overlapping clusters selected before the algorithm is run.

The k-Means algorithm works thus: given a set of d-dimensional training input vectors { x1, x2,.., xn }, the k-Means clustering algorithm partitions the n training examples into k sets of data points or clusters S = {S1, S2, …, Sk}, where k≤n, such that the within cluster sum of squares is minimised.

Generic k-means clustering Algorithms:

- Decide on the number of clusters, k.
- Initialize the k cluster centroids
- Assign the n data points to the nearest clusters.
- Update the centroid of each cluster using the data points therein.
- Repeat steps 3 and 4 until the changes in positions of centroids are zero

2.2 Decision Tree

It is a well known classification method that takes the form of tree structure and it is usually made up of:

- Testing node which holds the data for testing the condition
- Start node is the parent and usually top most node.
- Terminal node (leaf node): is the predicted class label
- Branches: represents results of a test made on an attribute.

Decision tree can be built using different methods, the first method developed was ID3 (Interactive, Dichotomiser) which later metamorphosed into C4.5 classifier. J48 classifier is an improved version of C4.5 decision tree classifier and has become a popular decision tree classifier. Classification and Regression Trees (CART) was later developed to handle binary trees. Thus, ID3, J48 and CART are basic methods of decision tree classification (Aman and Suruchi, 2011 [1]).

Decision trees are powerful and popular for both classification and prediction.

Decision tree algorithm proposed by Jiawei is presents further [4].

Algorithm

Parameters
Dataset T and its fields
Set of Attributes A
Selection Technique for the Attribute

Result
Tree Classifier

Procedure
1. A node is Created (call it E)
2. Check if all records R is in one group G and write node E as the last node in the that Group G
3. If $A = 0$(no attribute)

4. then write E as the last node
5. Use Selection technique for attributes on (R, A) to get the
   Best splitting condition
6. Write the condition on node E
7. Check if attribute is discrete and allows multiway split then
   It is not strictly binary tree
8. For all output O from splitting condition, divide the records and build the tree
9. Assign R $Set\ of\ all\ records\ in\ output\ O_0 =$
10. If $R_o = 0$then
11. Node E is attached with a leaf labelled with majority class R
12. Otherwise node E is attached with node obtained from
    Generate Decision Tree (R_o, A)
13. Next
14. Write E

3 Model specification and results

3.1 Input data

The survey is conducted on a representative sample of Romanian teachers. After the statistical processing of the collected data and after analysing the results derived from data processing, one shall establish the requirements for implementing the training programme for teachers in use of ICT in teaching. The research is also one of the landmarks which shall be used to assess the efficiency and impact of the training programme.

The research was conducted on a sample of 1,400 teachers. The extent of the sample satisfies one of the essential conditions for obtaining reliable results, namely the use of relatively large samples, which gives it representation at national level.

The method used to select study participants was random stratified sampling method, a very widely used method, leading to obtaining a sample under suitable conditions in terms of time and cost, relevant for the analysis of various groups of the surveyed population.

When determining the relevant population, one has taken into account to establish the assembly of individuals or organisations this research focuses on, and which its findings will be reflected on. Depending on the school level where the teachers in the target group work, the sample includes:

- Primary education teachers;
- Lower secondary school teachers;
- Higher secondary school teachers;
- Vocational education teachers;
- Higher education teachers.

The variables and indicators we have started from in drafting this instrument for data collection are:

- independent variables are identified by multiple choice questions (each with one possible answer):
  - identifying the respondent (question type: Information about respondents): the questionnaire enhances respondent's privacy; the respondent is identified by his/her name and surname initials;
  - occupation (type of institution) and experience (seniority): type of the institution where teachers carry out their professional activity;
  - environment (rural/ urban) where teachers carry out their professional activity;
  - gender of the surveyed teachers.

- **dependent variables** seek to test teachers' attitude towards their own professional status and their attitude towards continuous training, based on the following elements:
  - attitude towards their own professional status:

- teachers' satisfaction shown about their professional environment;
- perception on their own professional skills;
- attitude towards lifelong learning:
- teachers' conception of lifelong learning;
- frequency of accessing training courses;
- share of reasons for not participating in training;
- share of accessing training courses content;
- share of teachers' expectations towards future training areas;
- share of the need to develop professional skills;
- teachers' views on the expected features of training courses;
- teachers' views on the benefit of lifelong learning.

The extent to which respondents feel they need training to acquire and strengthen skills, in such areas as:

- knowledge and appropriate use of theoretical concepts related to ICT;
- designing educational content and assessment using specific ICT tools;
- knowledge and use of ICT tools;
- knowing and implementing teaching strategies enabling the effective use of ICT tools and teaching resources in the educational process;
- using basic functionalities specific to word processing programs;
- using basic functionalities specific to spreadsheet programs;
- using basic functionalities specific to presentations editing programs;
- using search programs to access the information available in the virtual environment and communicating by e-mail;
- knowing and getting acquainted with the operating mode of ICT tools providing facilities for converting teaching materials in accessible formats for various types of deficiencies experienced by disabled students.

These are multiple-choice questions, ranging from "to a very great extent" (value 5) up to "to a very small extent" (value 1).

3.2 Model definition

Clustering is a technique for extracting information from unlabelled data. Data Clustering is unsupervised and statistical data analysis technique. It is used to classify the same data into a homogeneous group. It is used to operate on a large data-set to discover hidden pattern and relationship helps to make decision quickly and efficiently. In a word, Cluster analysis is used to segment a large set of data into subsets called clusters.

Our model is built in several steps. In the first step, pre-processing data step, input data are normalized in order to be ready for processing. Then, we use divisive hierarchical clustering in order to obtain the optimum number of clusters.

Normalized input data and the optimum number of clusters are used by k-means algorithm. This algorithm is one of the simplest unsupervised learning algorithms that solve the clustering problem. The procedure follows a simple and easy way to classify a given data set through a certain number of clusters (assume k clusters) fixed apriori. The main idea is to define k centers, one for each cluster. These centers should be placed in a cunning way because of different location causes different result. So, the better choice is to place them as much as possible far away from each other. The next step is to take each point belonging to a given data set and associate it to the nearest center.

3.3 Model implementation

We chose RapidMiner (RM) for implementation of our model [6]. The main reasons which recommend RapidMiner for our model implementation are:

- Is one of the most powerful open-source systems for data mining.
- It includes a large collection of modular operators for design and processing of complex data mining problems
- Knowledge and data miner processes are represented by means of tree-operators. The leaves of the tree correspond to the simplest steps from the modelled process; the interior nods correspond to the abstract steps and the root to the whole process.
- For each operator are defined the input and output data and many settings parameters.
- All RapidMiner processes are described using XML
- It has a user friendly interface.
- It supports a flexible arrangement/rearrangement of operators
- It allows data import from a lot of formats (Excel, CSV, XML, Access, AML, ARFF, XRFF, SPSS, Stata, Sparse, DBase, C4.5, etc.)
- Offers many types of output data visualization thereby proving a easier understanding and interpretation of the results.

For the hierarchical clustering we use TopDownClustering operator. KMeans operator solve the non-hierarchical clustering task offering as output data the clusters. The representation of the clustering solution as a decision tree is realized using DecisionTree operator.
The implemented processes and the practical results are presented in the next section.

3.4 Practical results

First process is used in order to obtain the optimum number of clusters. It use an import operator named ReadExcel which reads an ExampleSet from the specified Excel file. The NominalToNumerical operator is used for pre-processing data named which changes the type of selected non-numeric attributes to a numeric type.

The TopDownClustering operator is used for hierarchical clustering and performs top down clustering by applying the inner flat clustering scheme recursively. Top down clustering is a strategy of hierarchical clustering. The result of this operator is a hierarchical cluster model.
The chains of the process is presented in figure 1.

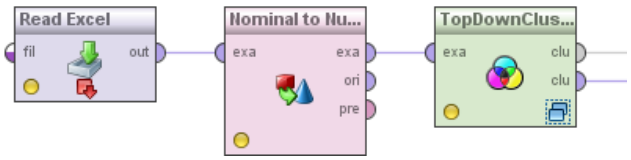

Figure 1. Top-down clustering

Second process is a combination of clustering and decision tree and is used to obtain the representation of the model. The chain of the process use KMeans operator which performs clustering using the *k-means* algorithm. This operator contains a parameter which specifies the number of clusters to form. The input value for this parameter is the value obtained in the first process.

The ChangeAttributeRole operator is used to change the role of one or more attributes. The Role of an attribute reflects the part played by that attribute in an ExampleSet. Changing the role of an attribute may change the part played by that attribute in a process. One attribute can have exactly one role. The target role for out attribute is label.

The final operator in our chain is DecisionTree. This operator generates a decision tree for classification of both nominal and numerical data. A decision tree is a tree-like graph or model. It is more like an inverted tree because it has its root at the top and it grows downwards. This representation of the data has the advantage compared with other approaches of being meaningful and easy to interpret.

The chains of the process is presented in figure 2.

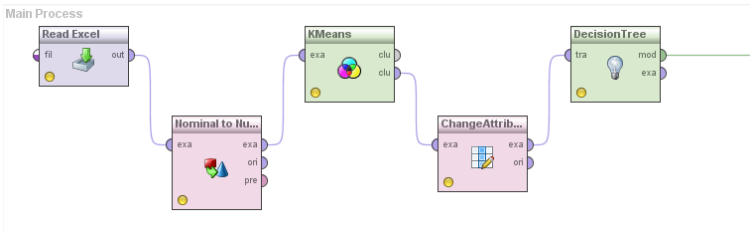

Figure 2. Non-hierarchical process

The clusters obtained using the process presented in figure 2 are show in figure 3.

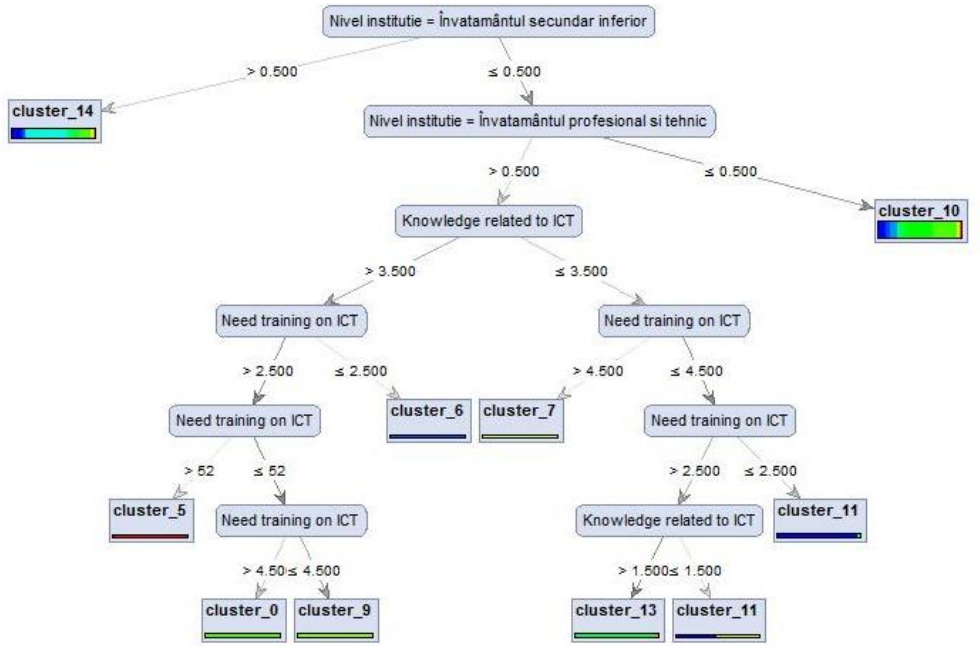

Figure 3. Decision Tree

These clusters help in the selection process in order to create the study groups and to adjust the level courses for each group.

It is suggested that, in establishing the course topics, one should take into consideration such objectives as: focusing on the learner; distributed resources by integrating electronic libraries and multimedia materials; open-minded, reusable learning objects making up adaptive training routes, virtual learning environments; computer assisted collaborative learning, social e-learning, social networking, asynchronous learning networks; simulations, educational computer games.

By attending the course, trainees should be shaped such attitudes as: the principle of equitable access of all students to information resources and technology; the adaptability to the information technology (development) needs and requirements; the use of interactive/ modern teaching methods; bringing virtual environment in the classroom space.

4 Conclusions

The course is recommended to provide information on: the use of ICT tools to streamline both teachers' own activity and the activity of their students; the information features (classification by source, validity, memory consumption, and the eventuality of changing and exchanging information), stressing that there should not be lost sight of, on one hand, the beneficiary-teacher, as a training participant and, on the other hand, the beneficiary-pupil/student, as the beneficiary of educational services offered by the trainees.

In designing the course topics, it is recommended to take into account that, at the end of the course, all participants should be able to use information technologies in their lessons, create digital resources to use in class, and use ICT technology for information and documentation.

In establishing the training methodology and strategies, one must take into account that most of the Romanian teachers are not sufficiently skilled to be enabled to know and get acquainted with the operating mode of ICT tools providing facilities for converting teaching materials in accessible formats for various types of deficiencies experienced by disabled students; to use basic functionalities specific to presentation editing programs; to use search programs to access the information available in the virtual environment and communicate by electronic mail; to know and use ICT tools; to use basic functionalities specific to word processing programs; to use basic functionalities specific to spreadsheet programs.

The course should be designed to facilitate access to technology and information channels and meet teachers' needs, such as: using search programmes to access the information available in the virtual environment and communicating by electronic mail, getting acquainted with digital devices and resources, as educational tools (both for learning and personal development.

Within this training programme, it is recommended to aim at: fostering attendees' creativity, structured thinking ability and interactivity. This training programme should also facilitate: collecting, analysing and interpreting data and information; taking into consideration individual differences and learning progress; use and maintenance of specialized educational software. Some of the key concepts within the course are recommended to be information, efficiency and applicability.

At the end of the training programme, the teacher's psychosocial portrait, as a training participant, should have the following features: to know when, how and where to use technology in his/her lesson, the teacher must know the basic hardware and software operations, and web resources suitable to the subject they teach, he/she should develop ICT based learning environments - to search, analyse and assess information - he/she should creatively and effectively use ICT tools.

The surveyed teachers consider they need training to acquire/ strengthen skills in using ICT tools to streamline both their own activity and the activity of their students. It follows that the approach of the course must be focused, on one hand, on the beneficiary-teacher, as a training participant and, on the other hand, on the beneficiary-pupil/ student, as the beneficiary of educational services offered by the trainees. The recommendation resulted from this situation is not to lose sight of the indirect recipient of the courses offered by this project: the pupil/ student.

This information, drawn from the study, allows us to establish as potential targets, which shall turn into skills, several needs clearly highlighted by the surveyed teachers: frequent use of information technologies in lessons, creating digital resources for classes, use of ICT for information and documentation.

References

[1] Aman Kumar Sharma, Suruchi Sahni, "A Comparative Study of Classification Algorithms for Spam Email Data Analysis", IJCSE, Vol. 3, No. 5, 2011, pp. 1890-1895

[2] A.K. Jain, M.N. Murty and P.J. Flynn,(1999)."Data Clustering: A Review". ACM Computing Surveys, Vol. 31, No. 3.

[3] Jean Yan,(2013). "Big Data, Bigger Opportunities- Data.gov's roles: Promote, lead, contribute, and collaborate in the era of big data". Retrieved from http://www.meritalk.com/pdfs/bdx/bdxwhitepaper-090413.pdf on 14 July 2015.

[4] Jiawei H., Micheline K., and Jian P. (2011)"Data mining: Concept and Techniques" 3rd edition, Elsevier,
[5] Vaishali R. Patel1 and Rupa G. Mehta, (2011). "Impact of Outlier Removal and Normalization Approach in Modified k-Means Clustering Algorithm". IJCSI International Journal of Computer Science Issues, Vol. 8, Issue 5, No 2, September 2011 ISSN (Online): 1694-0814
[6] I. Mierswa, M. Wurst, R. Klinkenberg, M. Scholz, T. Euler, Yale (now: RapidMiner): Rapid Prototyping for Complex Data Mining Tasks, Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2006), 2006.

Daniel Hunyadi
"Lucian Blaga" University of Sibiu
Department of Mathematics and Informatics
5-7 Dr. Ratiu Street 550012
Romania
E-mail: daniel.hunyadi@ulbsibiu.ro

Daniel Mara
"Lucian Blaga" University of Sibiu
Department of Private Law and Educational Siences
Romania
E-mail: danielmara11@yahoo.com