Active Clustering based Classification for Cost Effective Prediction in few Labeled Data Problem

Gábor SZŰCS, Zsuzsanna HENK
Department of Telecommunications and Media Informatics, BME, Hungary
szucs@tmit.bme.hu

In many data mining problems related to business it is hard to obtain labeled instances. When the labeled data set is not large enough the classifiers often perform poor results. Nevertheless semi-supervised learning algorithms, e.g. clustering based classification can learn from both labeled and unlabeled instances. We have planned and implemented a semi-supervised learning technique by combining the clustering based classification system with active learning. Our active clustering based classification method first clusters both the labeled and unlabeled data with the guidance of labeled instances, then queries the label of the most informative instances in an active learning cycle and after that classifies the data set. At cost benefit analysis comparing the results of our system with the supervised learning and clustering based classification it can be concluded that our solution saves the largest cost.
Keywords: *Active Learning; Data Mining; Semi-Supervised Learning; Clustering Based Classification; Cost Benefit Analysis*

1 Introduction

The information management process in business is concerned with the collection, collation, and use of customer data and information from all customer contact points. The key material elements of the information management process are the data repository, which provides a corporate memory of customers; customer analytics tools; front office, which supports many activities involved in interfacing directly with customers; and back office applications with managing internal operations, administration, and supplier relationships [9]. In information management process the companies use decision support tools, like expert systems, e.g. fuzzy expert system [5], simulation software, analytical tools, e.g. CRM, etc.

In customer relationship management (CRM) [6] the gathering information process determines the quality of the data, and in order to give appropriate marketing responses large amount of customer information is needed as well. CRM is related to long term success in the market [4] and in analytical CRM the data mining methods deal with only data, so the business problems should be formulated by data. The most popular data mining task is the classification, which is very useful from economic point of view, because prediction can be solved for a new instance based on labeled instances, where the class of the instance is known.

In large part of the real-world business problems the data is incomplete and Expected Maximization algorithm – as well as for the particular case of unsupervised statistical learning – is able to solve this problem [10].

In another large part of real-world data mining problems related to business it is hard to obtain labeled instances. The obtaining can be expensive, time-consuming or maybe the required data is not available; in these cases there is only a small labeled data set. The supervised learning methods are trained with labeled data, and only large amount of data leads to satisfactory learning, so there is a challenge to solve this contrast.

In this paper we present a new method to solve this problem by using a semi-supervised learning technique and applying active learning method together. Both approaches address the same problem just from different point of view. Semi-supervised learning algorithms can learn from not only labelled but unlabeled data as well, thus they can be used to improve the performance when there is only a small amount of labelled instances. On the other hand, active learning algorithms can choose the labelled instances

which from they learn and this way they achieve more accurate results with fewer instances. Our method combines these two approaches; we use a clustering based classification semi-supervised technique in a way that we determine the small labelled data set for the algorithm by asking the most informative instances' labels.

2 Related Works

Traditional machine learning and data mining algorithms predict future data using statistical models that are trained on previously collected training data. Semi-supervised classification [1][7][12]addresses the problem that the labeled data may be too few to build a good classifier, by making use of a large amount of unlabeled data and a small amount of labeled data. There is another new trend in data mining, the transfer learning [7], which allows the domains, tasks, and distributions used in training and testing to be different, but in this paper this trend is not considered, because we have dealt with only common domains and tasks in business problems.

At problems of small size of labeled instances there are two possible approaches for the classifying: inductive and transductive approach [3]. The inductive approach uses the labeled data to train a supervised learning algorithm, and then it predicts labels for all of the unlabeled instances. However, the supervised learning algorithm will only have very few labeled instances to use as a basis for building a predictive model. Transduction has the advantage of being able to consider all of the instances, not just the labeled data, while performing the labeling task. Transductive algorithm would label the unlabeled instances according to the clusters to which they naturally belong. An advantage of transduction is that it may be able to predict better with fewer labeled instances; but one disadvantage of transduction is that it builds no predictive model. If a previously unknown instance is added to the set, the entire transductive algorithm would need to be repeated with all data in order to predict a label.

The clustering based classification (CBC) [13] belongs to the transductive approach. CBC algorithm uses training data, including both the labeled and unlabeled data, which is first clustered with the guidance of the labeled data. Some of unlabeled data samples are then labeled based on the clusters obtained. The ratio of these samples in all unlabeled data can be changed by a *p* parameter. At the end of the whole method a discriminative classifiers can be trained with the expanded labeled dataset.

We have also used transductive approach, furthermore we would like to compare the semi-supervised classification (classifiers using both labeled and unlabeled training data samples) and our semi-supervised active learning approach, therefore the same training set and test set have been required. For this reason the whole data set has been divided into two subsets in the beginning: (i) training set with only very few labeled and many unlabeled data, and (ii) unchangeable test set, which cannot be used for active learning. The unchangeable test set has been the same for all algorithms in the comparison.

3 Clustering Based Classification by Modified K-Means Algorithm

For the solution of the problem mentioned above our idea was to build an intelligent system, which knows labels of only some random instance in the beginning, then enlarges the labeled instance set cyclical by active learning technique – which selects the most useful instances for the learning algorithm – until the set reaches the appropriate size. A clustering based classification has been used for building the classifier, and combination of this technique with active learning is the contribution of this paper.

At clustering based classification (without active learning) the training set consists of labeled and unlabeled instances as well, where the ratio of unlabeled ones is typically larger as mentioned earlier. A clustering algorithm creates clusters, and then the unlabeled instances or parts of them become labeled according to clusters. After clustering phase the supervised learning will use all the labeled instances (originally and new ones based on clusters). The classifier will build a

model based on extended labeled data set, which will be tested on unlearned (unknown) test set.

For clustering we have modified the original k-means algorithm. Our modified k-means algorithm begins by selection of labeled instances, and calculates centers of classes based on these (center coordinates are averages of coordinates of instances). These class centers will be the initial cluster centers, thus we hope that later during the process each cluster will represent a class. The clustering will start based on these initial cluster centers, and the labeled and unlabeled instances will be clustered during the two-phase iterative process. The clustering algorithm iterates until the results remain unchanged in two consecutive cycles. The algorithm of modified k-means can be seen below.

1. Number of clusters let equal to number of classes (k).
2. For each $i \in \{1, \ldots, k\}$, set c_i to be the center of mass of all labeled instances in i^{th} class, as can be seen in Equation (1), where ${}_mc_i$ and ${}_mx$ is the m^{th} coordinate of the c_i point and x point respectively, and $LClass_i$ is the set of initial labeled instances in the i^{th} class.

$$ {}_m c_i = \frac{\sum_{x \in LClass_i} {}_m x}{|LClass_i|} \qquad (1) $$

3. For each $i \in \{1, \ldots, k\}$, set the cluster C_i to be the set of points in X that are closer to c_i than they are to c_j for all $j \neq i$.
4. For each $i \in \{1, \ldots, k\}$, set c_i to be the center of mass of all points in C_i,, as can be seen in (2), where ${}_mc_i$ and ${}_mx$ is the m^{th} coordinate of the c_i point and x point respectively.

$$ {}_m c_i = \frac{\sum_{x \in C_i} {}_m x}{|C_i|} \qquad (2) $$

5. Repeat Steps 3 and 4 until C no longer changes.
6. Set the label of all unlabeled instance based on label of nearest center.

4 Active Clustering Based Classification

The active learning [11] can help with selecting the most useful instances into the small instances set for the learning system. In our solution in each active learning cycle a clustering phase can be found with modified k-means algorithm and labeling the instances as can be seen in Fig 1. The clustering phase is modified k-means algorithm and it is embedded in the iterative active learning process, because we use active learning indirectly for the clustering. In this cycle the clustering algorithm with constraint of labeled instances creates clusters of the labeled and unlabeled data set, and based on the results the algorithm inscribes the labels of unlabeled instances.

Figure 1 shows that the active learning phase will query the ground truth label of the instances whose label has the largest uncertainty (i.e. in our solution the largest distance between the instance and center of its cluster). Afterwards instances – in the answer – with label information will be added to labeled set, and then the clustering phase will be executed again. This will be continued until the size of the labeled data set reaches the appropriate size, which is equal to maximal available labeled set at simple clustering based classification.

In each active learning cycle the number of queries may be larger than 1, in our work this was equal to number of the classes, N_{class}. These queries come from different classes, so in each class there is a query, which request the label of most uncertain instance in that cluster (as representing the class). This does not definitely mean, that at the final end the number of instances will be equal in all classes, because the ground truth labels at the answer are not sure uniform distributed.

We have compared the effects of only one query and N_{class} queries in a cycle and our experiences have shown that the latter has been better. The theoretical reason of this de-

cision is based on the concept that in each cycle we would like to increase the goodness of each cluster corresponding to a class. Thus it cannot occur that the algorithm gets information from only one or few class, while the others will not develop.

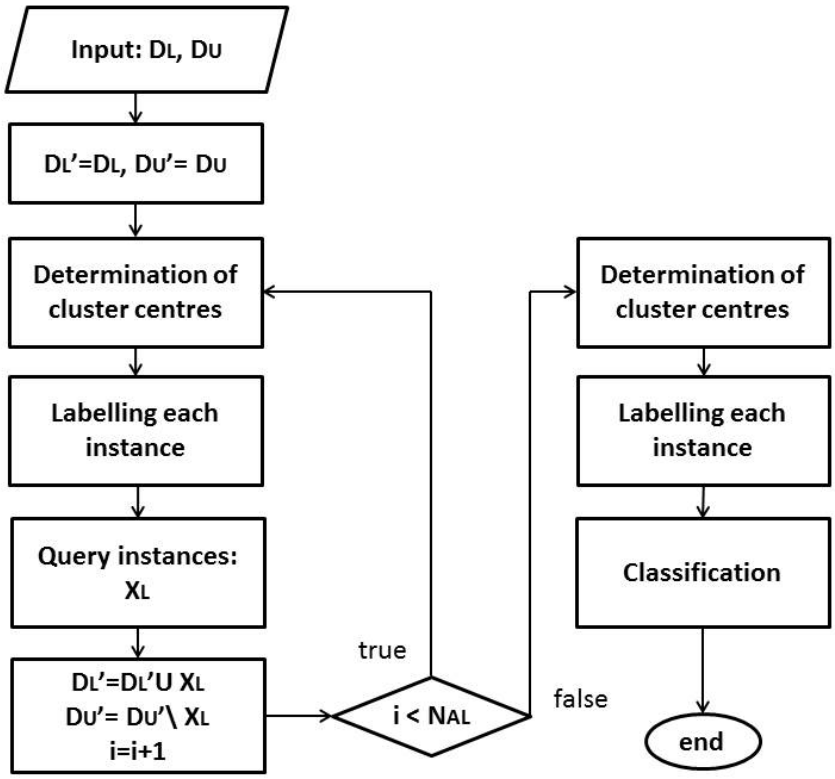

Fig. 1. Block schema of active clustering based classification

The number of cycles of iterative active learning process can be determined by the following: $N_{AL}=[(N_{endL}-N_{initL}) / N_{class}]+1$, where N_{initL} is the number of labeled instances in the beginning, N_{endL} is the number of labeled instances at the end, and the square brackets represent the down rounded value (integer).

So the active learning is solved by N_{AL} cycles. After the last query the new information will be used at an additional clustering process in order to get the best predictive labels. At the end of the clustering process the predictive label of an unlabeled instance will be the label of nearest center. Afterwards the algorithm selects the part (best percent *p*) of predictive labeled instances that/which possess the most certain label, and these will be the training set of supervised learning. This selection step needs to try to avoid the inaccurate classification because of very uncertain predictive labels. There are more methods for classifying the unlabeled instances, like linear and non-linear SVM [2], decision tree, neural networks, but at this paper we have used k-NN (k nearest neighbor) classification algorithm because of zero training time. The only modification from original algorithm is that our training set contains ground truth labeled and predictive labeled instances as well.

5 Testing

For possibility to comparison and to measure the improvement we have solved a problem by simple supervised learning, by semi-supervised learning (clustering based classification), and by our method, i.e. active clustering based classification. Some parts of analysis we have executed by SPSS (Statistical Package for the Social Sciences) Statistics 17.0, and by RapidMiner 5.2.

The quantity of the help of active learning is tested in a data set, and in order to see the differences the results of the simple classification, the clustering based classification and the combined method they are compared.

The problem has been a single-label, multi-class classification task with 26 classes. The data can be found in a website [3], the total set consists of 20 000 instances and each instance has 16 numerical attributes. The data set is balanced, i.e. the number of instances is approximately equal in each class, and there are no missing values.

We have investigated different classification algorithms: k-NN, SVM, and Naïve Bayes, and in the investigated data set the conditional reachable accuracies of them are follows:

$a_{k\text{-}NN} = 95.59\%$

$a_{SVM} = 84.16\%$.

$a_{NB} = 64.27\%$.

Conditional reachable accuracy means the accuracy with very large training set. For this data set the k-NN classifier has been the most accurate model, so in the analysis described below only this classifier is used. The analysis concerns to a small labeled set instead of large one presented above. The size of the labeled set has been equal in three investigated method: (i) in simple classification, (ii) in clustering based classification in the beginning of the clustering, and (iii) in active clustering based classification at the end of active cycles. The comparison is based on these conditions.

Expectedly the accuracy will be smaller than in large labeled data set, since the classifier can use fewer instances for building model. The aim is to increase this smaller accuracy by other methods (by clustering based classification and by our method).

The training set is generally 50-80% of the whole data set and the rest is the test set. This means that the training set would be 15000 and the test set would be 5000 in our selected data set at case of 3:1 division ratio. But the goal of this paper is the investigation of very few known data, so let the number of labeled data is 5 in each class (5*26=130, which is 0.87% of 15000).

We have used the same method and similar parameters for training and testing of simple classification (k-NN), and the accuracy (with only 130 training data) has been much smaller, than 95.59% as can be seen at 5000 test data:

$a_{k\text{-}NN} = 27.48\%$

At clustering based classification (CBC) the number of labeled data in the training set has been 130 again, the size of unlabeled training set has been 15000-130=14870, and the number of instances in test set has been 5000. We have tuned the parameter *p*, and we have found that approximately 50% has been the best in the investigated data set. In these circumstances the accuracy has been improved:

$a_{CBC} = 55.04\%$.

The third method has been the active clustering based classification (ABC), where the size of the unlabeled training set and test has been 14870 and 5000, respectively. The initial number of labeled data in the beginning of the active learning phase has been 26 and 130 at the end of the process. The parameter *p* has been also 50%. Since the selection of new instances for query the label is not random, therefore the accuracy has been improved again: $a_{ABC} = 60.27\%$.

In this example the number of learning cycles has been 5, which has been not enough to present a large improvement, but the increase can be seen.

The learning curve of the active clustering based classification can be seen in Fig. 2, where the accuracy is presented as the function of labeled data size.

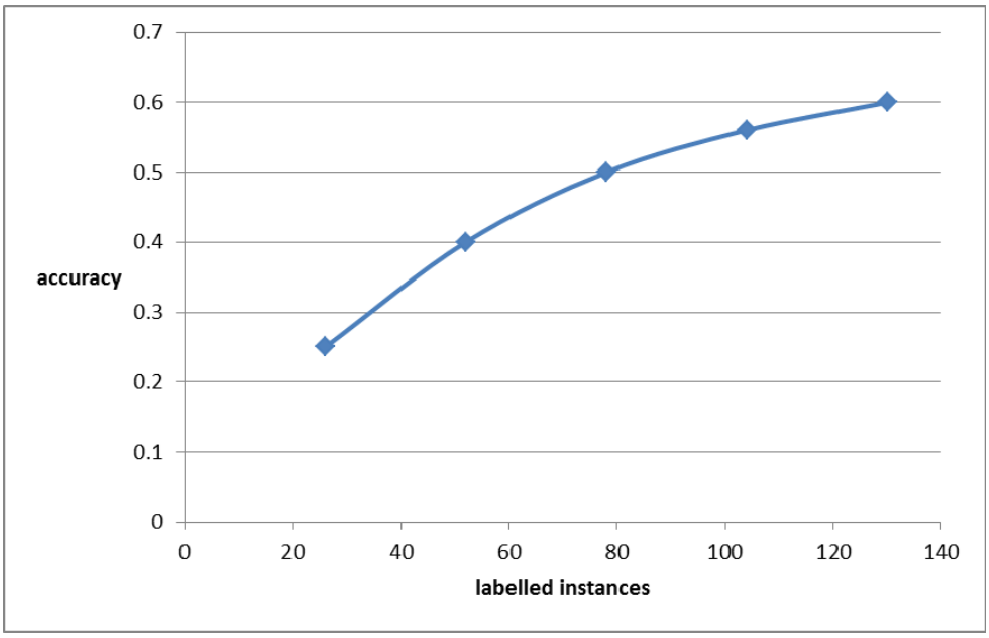

Fig. 2. Learning curve of the active clustering based classification

Summarizing the results the advantage of the active clustering based classification is presented, since using only 0.87% of maximal 15 000 training data set the accuracy is larger than 60% at the selected data set.

6 Cost Benefit Analysis

The data mining can help the enterprises with increase of competitiveness. The success depends on the capability for the gathering appropriate information and the best possible utilization of them. The information gathering is important for more reasons. One the one hand in the big data area huge amount of data can be accessed and only adequate information can be utilized among them; on the other hand the profit can be increased by reduction of costs, and this can be achieved by data mining.

In this session we investigate the possibility of decreasing the costs of data mining in business analytics, since in many problems the data can be accessed easily, however only unlabeled data are free and process of labeling has cost. At this kind of problem the aim of a classification task is the successful problem solving (satisfactory/ sufficient accurate classification) by the possible least labeling processes. This has been the goal at the elaboration of active clustering based classification.

How can be measured the costs, saves, profit in a data mining problem solution? The *cost benefit analysis* is able to measure these values, which is developed for projects and this is capable to present the best alternative among different ones.

The data mining can help in such problems, where the labeled data are not available, but obtaining of each labeled instance has cost, thus it should be decided that how many instances are required, and which instances should be queried for obtaining the label.

Let us suppose that the income and expenses come only at zero time. The expenses need for labeling process and the correctly classified instances will give the income. Let l be the cost of the labeling process of an instance and let us denote the future income of a correctly classified instance by b. Furthermore we suppose that the incorrectly classified instances are free of charge (there is neither income nor expenses). In order to keep the previously used data, let the size of the test set is 5000, so the task is to classify unknown 5000 instances.

At case of the simple supervised learning with 15000 (labeled) training and 5000 test set the profit is the following:

$profit = -15000 * l + 5000 * 0.9559 * b = -15000 * l + 4779.5 * b$

If this supervised learning possessed only

130 labeled instances in training set, then the profit is:
profit = -130 * l + 5000 * 0.2748 * b = -130 * l + 1374 * b
At case of clustering based classification with 130 labeled and 14870 unlabeled instances in training set the calculation of the profit can be seen:
profit = -130 * l + 5000 * 0.5504 * b = -130 * l + 2752 * b
At case of active clustering based classification with 130 labeled and 14870 unlabeled instances in training set the profit is:
profit = -130 * l + 5000 * 0.6027 * b = -130 * l + 3013.5 * b
We would like to know which alternatives are profitable, and which possesses the largest profit, but this depends on values l and b. In the comparison the ratio of l and b is enough to measure the difference among the alternatives, thus the profit can be drawn in a function of the ratio l/b.

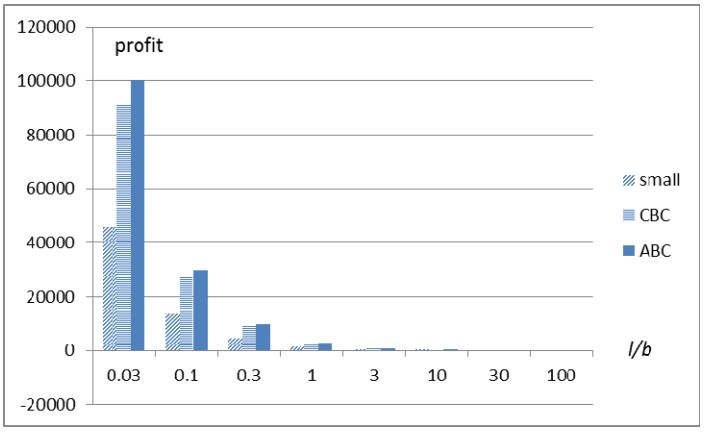

Fig. 3. Comparison of methods at small labeled data set

In the Fig. 3. comparison of methods can be seen at small labeled data set (with 130 labeled and 14870 unlabeled instances), where *small* is the simple supervised learning, *CBC* is the clustering based classification, and *ABC* is the active clustering based classification. The diagram shows that our ABC method can reach the largest profit in all cases.

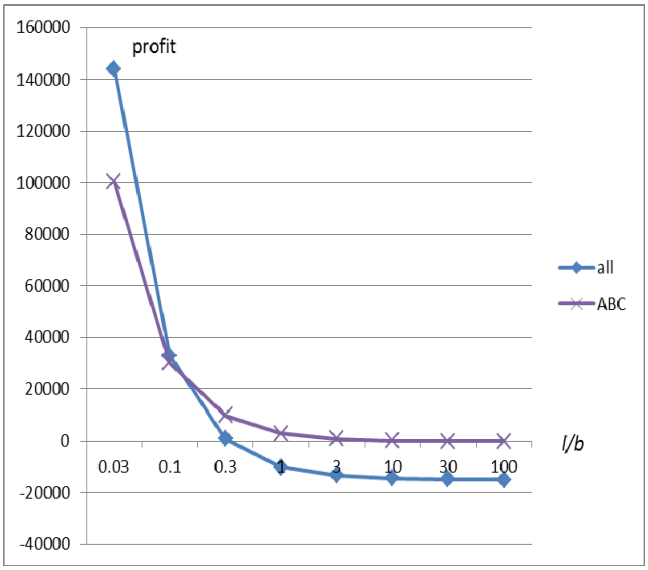

Fig. 4. Comparison of supervised and active clustering based classification

In the Figure 4 comparison of two methods can be seen, where *all* is the simple supervised learning with 15000 (labeled) training, and ABC is the active clustering based classification (with 130 labeled and 14870 unlabeled instances). The diagram shows that if

ratio l/b is larger than 0.119, then ABC gives larger profit (supervised learning tends to negative profit).

7 Conclusion

The aim of data mining project in business is to arrive at a business decision, to prepare for a decision, or to find the optimum. The decision helping by data mining will probably yield larger profit than the alternative without data mining tools. In many problems the obtaining of labeled instances are expensive, thus enterprises purchase only very small labeled data set, but our method is able to utilize this small set also. Our developed method, as part of semi-supervised learning algorithm family, is capable to gain information from not only the labeled, but unlabeled instances as well. Thus if the size of labeled data was small, then it is worth using this kind of method, since larger accuracy can be reached by extended (predicted) labeled data set, than only by ground truth labeled one.

The active learning deals with also problem of the small size of labeled data, thus this has been the next development phase in our construction. The key of this method is that the classifier can choose which instances may useful for the learning part of the method. Then this can probably reach larger accuracy with smaller labeled data set, since the labels of some instances may more informative than others.

By the combination of k-means clustering based classification and active learning we have constructed and implemented a new method. Based on a very small initial labeled data – which are available in the beginning – the method determines which instances are required for obtaining the label of them. The method increases the size of labeled set periodically until previously determined size. In each cycle a clustering method has been executed using the actual labeled data and after that the active learning part of the method queries the most uncertain instances. After the reaching the predetermined size the method calls the k-NN classification part.

At the cost benefit analysis we have compared the simple supervised classification, the semi-supervised classification (classifiers using both labeled and unlabeled training data samples) and our semi-supervised active learning approach at the same training set and test set. The results of the test show that accuracy of our active clustering based classification method is always better than others in small number of labeled data.

8 Further Development

Embedding of the clustering based classification in active learning cycle can be solved in two different ways. These ways depend on the final phase of an active learning cycle, which will be clustering or classification phase. In our solution the active learning cycle has finished after the clustering phase, but the other way can be investigated, and our future plan is to develop this.

If an active learning cycle would finish after the classification phase, then in a cycle the classification phase will directly follow the clustering phases. The training set of supervised learning consists of ground truth labeled set and the part of the predictive labeled set by clustering phase. The classifier would build a model based on this training set and would classify the unlabeled data set giving the probabilities of classification decision as well. The most uncertain instances would be queried, thus the size of the labeled data set would increase. This would be continued until the appropriate size of the labeled data, and implementation of this iterative process will be one of the further developments in future work.

References

[1] A. Blum and T. Mitchell, Combining Labeled and Unlabeled Data with Co-Training. *Proc. 11th Ann. Conf. Computational Learning Theory*, 1998, pp. 92-100.

[2] C. Cocianu and L. State, Kernel-based Methods for Learning Non-linear SVM. *Economic Computation and Economic Cybernetics Studies and Research, 47*(1), (2013), pp. 41-60.

[3] T. Joachims, Transductive Inference for Text Classification Using Support Vector

Machines. *Proc. 16th Int'l Conf. Machine Learning*, 1999, pp. 825-830.

[4] R. Korsakiene, The innovative approach to relationships with customers. *Journal of business economics and management, 10*(1), (2009), pp. 53-60.

[5] A. Mahmoudi, H. Shavandi, A Genetic Fuzzy Expert System To Optimize Profit Function. *Economic Computation and Economic Cybernetics Studies and Research, 46*(4), (2012), pp. 241-259.

[6] E. W. Ngai, L. Xiu, and D. C. K. Chau, Application of data mining techniques in customer relationship management: A literature review and classification. *Expert Systems with Applications, 36*(2), (2009), pp. 2592-2602.

[7] K. Nigam, A. K. McCallum, S. Thrun, and T. Mitchel, Text Classification from Labeled and Unlabeled Documents Using EM. *Machine Learning, 39*(2-3), (2000), pp. 103-134.

[8] S. J., Pan and Q. Yang, A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering, 22*(10), (2010), pp. 1345-1359.

[9] A. Payne and P. Frow, A strategic framework for customer relationship management. *Journal of marketing*, (2005), pp. 167-176.

[10] G. Ruxanda and I. Smeureanu, Unsupervised Learning with Expected Maximization Algorithm. *Economic Computation and Economic Cybernetics Studies and Research, 46*(1), (2012), pp. 17-44.

[11] B. Settles, *Active learning literature survey.* Computer Sciences Technical Report 1648, University of Wisconsin-Madison, 2010.

[12] D. J. Slate, UCI Machine Learning Repository: Letter Image Recognition Data - http://repository.seasr.org/Datasets/UCI/arff/letter.arff., 1991.

[13] H. J. Zeng, X. H. Wang, Z. Chen, H. Lu and W. Y. Ma, CBC: Clustering based text classification requiring minimal labeled data. *ICDM 2003, Third IEEE International Conference on Data Mining*, 2003, pp. 443-450.

[14] X. Zhu, *Semi-Supervised Learning Literature Survey.* Technical Report 1530, Univ. of Wisconsin-Madison, 2006.

Gábor SZŰCS was born in 1970 in Hungary. He has received MSc in Electrical Engineering from Budapest University of Technology and Economics (BME) in 1994. He is experienced in modeling and simulation, railway systems, traffic systems; he has received PhD degree in this field from BME in 2002. His further and currently research areas are data mining in business areas, multimedia mining, content based image retrieval, semantic search. He is associate professor at Department of Telecommunications and Media Informatics of BME. The number of his publications is more than 80. He is president of the Hungarian Simulation Society (EUROSIM). He has earned János Bolyai Research Scholarship of the Hungarian Academy of Science in 2008.

Zsuzsanna HENK has received her Bachelor Degree in Computer Engineering from Budapest University of Technology and Economics and Master Degree from Corvinus University of Budapest. Her main field of interest is data mining, business intelligence and analytics.