DISTRIBUTED LEARNING ALGORITHM BASED ON DATA REDUCTION

Ireneusz Czarnowski and Piotr Jędrzejowicz
Department of Information Systems, Gdynia Maritime University, Morska 83, 81-225 Gdynia, Poland

Keywords: Distributed data mining, Distributed learning classifiers, Data reduction, Agent-based approach.

Abstract: The paper presents an approach to learning classifiers from distributed data, based on a data reduction at a local level. In such case, the aim of data reduction is to obtain a compact representation of distributed data repositories, that include non-redundant information in the form of so-called prototypes. In the paper data reduction is carried out by simultaneously selecting instances and features, finally producing prototypes which do not have to be homogenous and can include different sets of features. From these prototypes the global classifier based on a feature voting is constructed. To evaluate and compare the proposed approach computational experiment was carried out. The experiment results indicate that data reduction at the local level and next merger of prototypes into the global classifier can produce very good classification results.

1 INTRODUCTION

Usually data mining algorithms base on the assumption that all the training data can be pooled together in a centralized data repository. In the real life there are, however, numerous cases where the data have to be physically distributed due to some constraints (for example, data privacy or others).

Applying traditional data mining tools to discover knowledge from distributed data sources may not be possible (Kargupta et al., 1999). In the real life it is often unrealistic or unfeasible to collect distributed data for centralized processing. The need to extract potentially useful patterns out of separated, distributed data sources created a new, important and challenging research area, known as the distributed data mining or knowledge discovery from multi-databases (Xiao-Feng Zhang et al., 2004).

Recently, several approaches to distributed classification have been proposed. In (Prodromidis et al., 2000) a meta-learning process was proposed as a learning tool for combining a set of locally learned classifiers into the global classifier. Meta-learning involves running, possibly in parallel, learning algorithms for each distributed database or set of data subsets from an original database, and than combining predictions from classifiers learned from the distributed sources by recursively learning ”combiner” and ”arbiter” models in a bottom-up tree manner. Generally, meta-learning methodologies view data distribution as a technical issue and treat distributed data sources as parts of a single database. It was pointed out in (Tsoumakas et al., 2004) that such an approach offers rather a narrow view of the distributed data mining, since the data distributions in different locations often are not identical, and is considered as sub-optimal heuristic.

Tsoumakas et al. (2004) proposed approach based on clustering local classifier models induced at physically distributed databases. This approach groups together classifiers with their similar behavior and with final indication of a classification model for each cluster that together guarantees a better results than the single global model. This approach belongs to the set of methods based on common methodology for distributed data mining known as a two-stage (Xiao-Feng Zhang et al., 2004).

Generally, two-stage methods base on extraction of prototypes from distributed data sources. The first stage involves the local data analysis, the second combines or aggregates the local results producing the global classifier.

Another two-stage approach to confront the discussed problem is to move all data from distributed repositories to a central location and to merge the data together for a global model building. Such approach escapes the sub-optimality problems of local models combination as it was pointed out by Tsoumakas et

Czarnowski I. and JÄŹdrzejowicz P. (2009).
DISTRIBUTED LEARNING ALGORITHM BASED ON DATA REDUCTION.
In *Proceedings of the International Conference on Agents and Artificial Intelligence*, pages 198-203
DOI: 10.5220/0001655401980203

al. (2004). However, moving all data into a centralized location can be limited by communication bandwidth among sites and may be too expensive. Selecting out of the distributed databases only the relevant data can eliminate or reduce the above restriction. Selection of relevant data at local sites can also speed up the data transfer for centralized learning and global knowledge extraction. Selection of relevant data is very often referred to as data reduction with the objective to find patterns, called prototypes, references vectors, or regularities within certain attributes (see e.g., Liu et al., 1998). Generally, the goal of data reduction approaches is to reduce the number of instances in each of the distributed data subsets without loss of extractable information, to enable either pooling the data together and using some mono-database mining tools or effectively applying meta-learning techniques. Learning models based on the reduced data sets combined later into a meta-model seems to be one of the most successful current approaches to distributed data mining (Stolfo et al., 1997). Learning classifiers on the reduced data sets and then combining them is computationally much more efficient then moving all distributed data sets into a centralized site for learning a global model.

The paper deals with a distributed learning classifier. The proposed approach involves two stages. At the local level the prototype selection from distributed data is carried out. Prototypes are selected by simultaneously reducing data set in two dimensions through selecting reference instances and removing irrelevant attributes. The prototype selection proposed in this paper is an extension of the approach introduced by Czarnowski and Jędrzejowicz (2008a), where instance reduction only was proposed. In the present approach the data reduction scheme is carried out independently at each site through applying an agent-based population search. Thus obtained prototypes do not have to be homogenous and can be based on different sets of of features. Next, at the second stage, the prototypes are merged at the global level and classifier models are combined to produce a meta-classifier called combiner.

The paper is organized as follows. Section 2 contains problem formulation and provides basic definitions. Section 3 explains the proposed agent-based population learning algorithm, that has been used for reducing distributed data sets, and provides details on how the combiner classifier is constructed. Section 4 contains results of the computational experiment carried out with a view to validate the proposed approach. Finally, the last section contains conclusions and suggestions for future research.

2 DEFINITIONS AND PROBLEM FORMULATION

The problem of learning from data can be formulated as follows: Given a data set D, a set of hypothesis H, a performance criterion P, the learning algorithm L outputs a hypothesis $h \in H$ that optimize P. In pattern classification application, h is a classifier. The data D consists of N training examples. Each example is described by a set of attributes A and is labeled with a class. The total number of attributes is equal to n. The goal of learning is to produce a hypothesis that optimizes the performance criterion (e.g. function of accuracy of classification, complexity of the hypothesis, classification cost or classification error).

In the distributed learning a data set D is distributed among K data sources $D_1,\ldots,D_K$, with $N_1,\ldots,N_K$ respectively, where $\sum_{i=1}^{K} N_i = N$ and where all attributes are presented at each location. In the distributed learning a set of constraints Z can be imposed on the learner. Such constraints may for example prohibit transferring data from separated sites to the central location, or impose a physical limit on the amount of information that can be moved, or impose other restrictions to preserve data privacy. The task of the distributed learner L_d is to output a hypothesis $h \in H$ optimizing P using operations allowed by Z.

In case of using prototypes as suggested in this paper, the data sources $D_1,\ldots,D_K$ are replaced by reduced subsets $S_1,\ldots,S_K$ of local patterns, which in general are heterogeneous. In this case $A_1,\ldots,A_K$ are sets of attributes from sites $1,\ldots,K$ respectively. However, it is possible that some attributes can be shared across more then one reduced data set S_i, where $i = 1,\ldots,K$.

Thus, the goal of data reduction is to find subset S_i from given D_i by reducing of number of examples or/and feature selection and with retaining essentially extractable knowledge and preserving the quality of data mining results. Let the cardinality of the reduced data set S_i be denoted as $card(S_i)$. Then the following inequality holds $card(S_i) < card(D_i)$. Similarly, when the cardinality of the set of attributes A_i is denoted as $card(A_i)$, the following inequality holds $card(A_i) < card(A)$. Ideally $card(A_i) \ll card(A)$. It is expected that the data reduction results in data compression. The instance reduction compression rate C_D is defined as $C_D = \frac{N}{\sum_{k=1}^{K} card(S_k)}$. The attribute selection compression rate C_A is defined as $C_A = \frac{n}{card(\bigcup_{i=1}^{K} A_i)}$. Overall, data reduction guarantees total compression equal to $C = C_D C_A$.

3 AN AGENT-BASED APPROACH TO LEARNING CLASSIFIER FROM DISTRIBUTED DATA

3.1 Main Features of the Proposed Approach

It is well known that instance reduction, feature selection and also learning classifier from distributed data are computationally difficult combinatorial problems (Dash and Liu, 1997; Rozsypal and Kubat, 2003).

Although a variety of data reduction methods have been so far proposed in the literature (see, for example Dash and Liu, 1997; Raman and Ioerger, 2003, Rozsypal and Kubat, 2003; Skalak, 1994; Vucetic and Obradovic, 2000), no single approach can be considered as superior nor guaranteeing satisfactory results in the process of learning classifiers.

To overcome some of the difficulties posed by computational complexity of the distributed data reduction problem it is proposed to apply the population-based approach with optimization procedures implemented as an asynchronous team of agents (A-Team). The A-Team concept was originally introduced by Talukdar et al. (1996). The design of the A-Team architecture was motivated by other architectures used for optimization including blackboard systems and genetic algorithms. Within the A-Team multiple agents achieve an implicit cooperation by sharing a population of solutions, also called individuals, to the problem to be solved. An A-Team can be also defined as a set of agents and a set of memories, forming a network in which every agent remains in a closed loop. All the agents can work asynchronously and in parallel. Agents cooperate to construct, find and improve solutions which are read from the shared, common memory.

In our case the shared memory is used to store a population of solutions to the data reduction problem. Each solution is represented by the set of prototypes i.e. by the compact representation of the data set from given local level. The team of agents is used to find the best solution at the local level and them the agent responsible for managing all stages of the data mining is activated.

3.2 Solution Representation

Population of solutions to data reduction problem consists of feasible solutions. A feasible solution *s*, corresponding to the set of selected data, is represented by a string consisting of numbers of selected reference instances and numbers of selected features. The first t numbers represent instance numbers from the reduced data set D_i, where t is determined by a number of clusters of potential reference instances. The value of t is calculated at the initial population generation phase, where at first, for each instance from original set, the value of its similarity coefficient, proposed by Czarnowski and Jędrzejowicz (2004), is calculated, and then instances with identical values of this coefficient are grouped into clusters. The second part of the string representing a feasible solution consists of numbers of the selected features. The minimum number of features is equal to one.

3.3 Agents Responsible for Data Reduction

Data reduction is carried out, in parallel, for each distributed data site. Data reduction, carried-out at a data site, is an independent process which can be seen as a part of the distributed data learning. Each data reduction subproblem is solved by two main types of agents. The first one - *optimizing agents*, are implementations of the improvement algorithms, each *optimizing agent* represents a single improvement algorithm. The second one, called the *solution manager*, is responsible for managing the population of solutions and updating individuals in the population. Each *solution manager* is also responsible for finding the best solution for the given learning classifier subproblem.

The *solution manager* manages the population of solutions, which at the initial phase is generated randomly and stored in the shared memory. When the initial population of solutions is generated the *solution manager* runs a procedure producing clusters of potential reference instances. Next the *solution manager* continues reading individuals (solutions) from the common memory and storing them back after attempted improvement until a stopping criterion is met. During this process the *solution manager* keeps sending single individuals (solutions) from the common memory to *optimizing agents*. Solutions forwarded to *optimizing agents* for improvement are randomly drawn by the *solution manager*. Each *optimizing agent* tries to improve quality of the received solutions and afterwards sends them back to the *solution manager*, which, in turn, updates common memory by replacing a randomly selected individual with the improved one.

To solve the data reduction problem four types of *optimizing agents* representing different improvement procedures, proposed earlier by Czarnowski and Jędrzejowicz (2008b) for the non-distributed case, have been implemented.

These procedures include: local search with tabu list for instance selection, simple local search for instance selection, local search with tabu list for feature selection and hybrid local search for instance and feature selection, where the both parts of the solution are modified with the identical probability equal to 0.5.

In each of the above cases the modified solution replaces the current one if it is evaluated as a better one. Evaluation of the solution is done by estimating classification accuracy of the classifier, which is created taking into account the instances and features indicated by the solution. In all cases the constructed classifier is based on the C 4.5 algorithm (Quinlan, 1993).

If, during the search, an agent successfully has improved the received solution then it stops and the improved solution is transmitted to the *solution manager*. Otherwise, agents stop searching for an improvement after having completed the prescribed number of iterations.

3.4 Agent Responsible for Managing the Process of Distributed Learning

The proposed approach deals with several data reduction subproblems solved in parallel. The process is managed by the *global manager*, which is activated as the first within the learning process. This agent is responsible for managing all stages of the data mining. At the first step the *global manager* reads the distributed data mining task that should be solved. Than *global manager* runs, in parallel, all subtasks, that correspond to independent learning classifiers problem.

When all the subtasks have been solved, solutions from the local level are used to obtain a global solution. Thus, the *global manager* merges local solutions and finally produces the global classifier, called also meta-classifier.

To compute the meta-classifier a combiner strategy based on voting mechanism has been applied. This combiner strategy is used to obtain a global classifier from the global set of prototypes. The global set of prototypes is created by integration of local level solutions representing heterogeneous sets of prototypes. To integrate local level solutions it has been decided to use the unanimous voting mechanism. Only features that were selected by data reduction algorithms from all distributed sites are retained and the global classifier is formed based on the C 4.5 algorithm.

4 COMPUTATIONAL EXPERIMENT RESULTS

To validate the proposed approach computational experiment has been carried out. The aim of the experiments was to evaluate to what extend the proposed approach could contribute towards increasing classification accuracy of the global classifier induced on the set of prototypes selected from autonomous distributed sites by applying an agent-based population learning algorithm. Classification accuracy of the global classifiers obtained using the set of prototypes has been compared with the results obtained by pooling together all instances from distributed databases, without data reduction, into the centralized database and with results obtained by pooling together instances selected from distributed databases based on the reduction of example space only. Generalization accuracy has been used as the performance criterion.

The experiment involved three data sets - *customer* (24000 instances, 36 attributes, 2 classes), *adult* (30162, 14, 2) and *waveform* (30000, 21, 2). For the first two datasets the best known and reported classification accuracies are respectively 75.53% and 84.46%. These results have been obtained from (Asuncion and Newman, 2007) and (”The European Network”, 2002). The reported computational experiment was based on the ten cross validation approach. At first the available datasets were randomly divided into the training and test sets in approximately 9/10 and 1/10 proportions. The second step involved the random partition of the previously generated training sets into the training subsets each representing a different dataset placed in a separate location. Next, each of the obtained datasets has been reduced using the agent-based population algorithm. The reduced subsets have been then used to compute the global classifier using the proposed combiner strategy. Such scheme was repeated ten times, using a different dataset partitions as the test set for each trial.

Computations have been run with the size of initial population set to 50. A number of repetitions for each improvement procedure was set to 100.

The above described experiment has been repeated four times for the four different partitions of the training set into a multi-database. The original data set was randomly partitioned into 2, 3, 4 and 5 multi-datasets of approximately similar size. The respective experiment results are shown in Table 1. The results cover two independent cases. In the first case only reference instance selection at the local level has been carried out, and next the global classifier has been computed based on the homogenous set of prototypes. In the second case full data reduction at the

Table 1: Average classification accuracy (%) obtained by the C 4.5 algorithm and its standard deviation.

Problem	number of distributed data sources			
	2	3	4	5
	Selection of reference instances at the local level only			
customer	68.45±0.98	70.40±0.76	74.67±2.12	75.21±0.7
adult	86.20 ±0.67	87.20±0.45	86.81±0.51	87.10±0.32
waveform	75.52±0.72	77.61±0.87	78.32±0.45	80.67±0.7
	Combiner strategy based on the feature voting			
customer	69.10 ±0.63	73.43 ±0.72	75.35 ±0.53	77.20 ±0.49
adult	88.90 ±0.41	87.45 ±0.31	91.13 ±0.23	91.58 ±0.41
waveform	80.12 ±1.03	82.46 ±0.98	85.04±0.73	83.84±0.64

local level has been carried out and the global classifier has been computed by the combiner strategy based on the feature voting.

Table 2: Compression ratio versus the number of distributed data sources.

Problem		number of distributed data sources			
		2	3	4	5
customer	C_D	192.9	124.1	100.5	88.2
	C_A	1.9	1.6	1.8	1.6
	C	357.9	196.0	177.3	140.4
adult	C_D	79.6	68.0	60.7	52.7
	C_A	1.2	1.4	1.3	1.2
	C	98.6	98.2	81.0	62.5
waveform	C_D	56.5	51.8	46.7	39.7
	C_A	1.3	1.4	1.2	1.3
	C	71.0	73.2	56.0	52.1

It should be noted that data reduction in two dimensions (selection of reference instances and feature selection) assures better results in comparison to data reduction only in one dimension i.e. instance dimension. The approach to learning classifier from distributed data, based on data reduction at the local level, produces reasonable to very good results. The data reduction at the local level resulted in both: a very good accuracy of classification, better then for "full dataset" (calculated though pooling at the global level all instances from local levels), and a very high data compression rate (see, for example, Table 2).

5 CONCLUSIONS

The paper presents an approach to learning classifiers from distributed data, based on a data reduction at the local level. At the global level the combiner classifier with feature selection through majority voting has been constructed and implemented. Computational experiment carried out has shown that the proposed approach can significantly increase classification accuracy as compared with learning classifiers using centralized data pool. An extensive compression of the dataset size at the global level and parallel computations at distributed locations are additional features increasing the efficiency of the approach.

Computational experiment results confirmed that the global classifier based on data reduction at the local level can produce very good results. However, the quality of results depends on the choice of strategy used for constructing the combiner.

Future work will focus on evaluating other combiner classifier strategies in terms of classification accuracy and computation costs.

ACKNOWLEDGEMENTS

This research has been supported by the Polish Ministry of Science and Higher Education with grant for years 2008-2010.

REFERENCES

Asuncion, A., Newman, D.J. (2007). UCI Machine Learning Repository (http://www.ics.uci.edu/ mlearn/MLRepository.html).

Irvine, CA: University of California, School of Information and Computer Science.

Czarnowski, I., Jędrzejowicz, P. (2004) An approach to instance reduction in supervised learning. In: Coenen F., Preece A. and Macintosh A. (Eds.), Research and Development in Intelligent Systems XX, Proc. of AI2003, the Twenty-third SGAI International Conference on Innovative Techniques and Applications of Artificial Intelligence, Springer-Verlag London Limited, 267-282.

Czarnowski, I., Jędrzejowicz, P., Wierzbowska, I. (2008a) An A-Team Approach to Learning Classifiers from Distributed Data Sources. In: Ngoc Thanh Nguyen, Geun Sik Jo, Robert J. Howlett,an Lakhmi C. Jain (Eds.), KES-AMSTA 2008, Lecture Notes in Computer Science, LNAI 4953, Springer-Verlag Berlin Heidelberg, 536-546

Czarnowski, I., Jędrzejowicz, P. (2008b) Data Reduction Algorithm for Machine Learning and Data Mining. In: Nguyen N.T. et al. (eds) IEA/AIE 2008, Lecture Notes in Computer Science, LNAI 5027, Springer-Verlag Berlin Heidelberg, 276-285.

Dash, M., & Liu H. (1997). Feature selection for classification. *Intelligence Data Analysis 1(3)*, 131-156.

Kargupta, H., Byung-Hoon Park, Daryl Hershberger, & Johnson, E. (1999). Collective Data Mining: A New Perspective Toward Distributed Data Analysis. In Kargupta H and Chan P (Eds.), *Advances in Distributed Data Mining.* AAAI/MIT Press, 133-184.

Liu, H., Lu, H., & Yao, J. (1998). Identifying Relevant Databases for Multidatabase Mining. In *Proceedings of Pacific-Asia Conference on Knowledge Discovery and Data Mining*, 210-221.

Prodromidis, A., Chan, P.K., & Stolfo, S.J. (2000). Meta-learning in Distributed Data Mining Systems: Issues and Approaches. In H. Kargupta and P. Chan (Eds.) *Advances in Distributed and Parallel Knowledge Discovery*, AAAI/MIT Press, Chapter 3.

Raman, B., & Ioerger, T.R. (2003). Enhancing learning using feature and example selection. *Journal of Machine Learning Research* (in press)

Rozsypal, A., & Kubat, M. (2003). Selecting Representative Examples and Attributes by a Genetic Algorithm. *Intelligent Data Analysis, 7(4)*, 291-304.

Quinlan, J.R. (1993). *C4.5: programs for machine learning*. Morgan Kaufmann, SanMateo, CA.

Skalak, D.B. (1994). Prototype and Feature Selection by Sampling and Random Mutation Hill Climbing Algorithm. *Procciding of the International Conference on Machine Learning*, 293-301.

Stolfo, S., Prodromidis, A.L., Tselepis, S., Lee, W., & Fan. D.W. (1997). JAM: Java Agents for Meta-Learning over Distributed Databases. In *Proceedings of the 3rd International Conference on Knowledge Discovery and Data Mining*, Newport Beach, CA, AAAI Press, 74-81.

Talukdar, S., Baerentzen, L., Gove, A., & P. de Souza (1996). Asynchronous Teams: Co-operation Schemes for Autonomous. Computer-Based Agents, *Technical Report EDRC 18-59-96*, Carnegie Mellon University, Pittsburgh.

The European Network of Excellence on Intelligence Technologies for Smart Adaptive Systems (EUNITE) - EUNITE World Competition in domain of Intelligent Technologies (2002). Accesed on 1 September 2002 from http://neuron.tuke.sk/competition2.

Tsoumakas, G., Angelis, L., & Vlahavas, I. (2004). Clustering Classifiers for Knowledge Discovery from Physical Distributed Database. *Data & Knowledge Engineering, 49(3)*, 223-242.

Xiao-Feng Zhang, Chank-Man Lam, & William K. Cheung (2004). Mining Local Data Sources For Learning Global Cluster Model Via Local Model Exchange. *IEEE Intelligence Informatics Bulletine, Vol. 4, No. 2.*

Vucetic, S., & Obradovic, Z. (2000). Performance Controlled Data Reduction for Knowledge Discovery in Distributed Databases, *Procciding of the Pacific-Asia Conference on Knowledge Discovery and Data Mining*, 29-39.