Automatic Categorization of Email into Folders by Ant Colony Decision Tree and Social Networks

Urszula Boryczka, Barbara Probierz and Jan Kozak

Abstract This paper presents a new approach to an automatic categorization of email messages into mailbox folders. The aim of this paper is to create an algorithm that would allow one to improve the classification of emails into folders by using solutions that have been applied in Ant Colony Decision Tree (ACDT). Additionally, elements of Social Network Analysis (SNA) were included in this algorithm. The new algorithm that is proposed here was tested on the publicly available Enron E-mail data set and all experiments were conducted on uncleaned data. For the purpose of comparing the results, additional tests were carried out by using selected classifiers which were generally available. The obtained results confirm that the proposed approach allows one to improve the accuracy with which new emails are assigned to particular folders based on an analysis of previous correspondence, even when uncleaned data sets are used.

Keywords Ant colony optimization · Social network analysis · Enron E-mail

1 Introduction

The history of email messages began almost half a century ago, i.e. when Louis Pouzin, Glenda Schroeder and Pat Crisman sent an email from one user to another in 1965. Unfortunately, this email service only made it possible to leave such a message for the other users of the same computer, whereas an email address had not been invented yet.

U. Boryczka · B. Probierz (✉) · J. Kozak
Institute of Computer Science, University of Silesia, Będzińska 39, 41–200
Sosnowiec, Poland
e-mail: barbara.probierz@us.edu.pl

U. Boryczka
e-mail: urszula.boryczka@us.edu.pl

J. Kozak
e-mail: jan.kozak@us.edu.pl

I. Czarnowski et al. (eds.), *Intelligent Decision Technologies 2016*,
Smart Innovation, Systems and Technologies 57,
DOI 10.1007/978-3-319-39627-9_7

It was not until 1971 that the American engineer and computer programmer Raymond S. Tomlinson came up with an idea that allowed one to send an email message from one computer to another. In order to separate the user name from the computer name, Raymond S. Tomlinson picked the @ symbol, which was only used occasionally at that time. On this basis, members of the Internet Engineering Task Force agreed on the standard syntax for email communication in 1973, i.e. "username@hostname", which is still used to this day.

The presented paper deals with research that is focused on creating decision tables in accordance with the authors' idea and testing classical algorithms by using these tables. The aim of this paper is to create an algorithm that would allow one to improve the classification of emails into folders (so. E-mail Foldering Problem) by using solutions that have been applied in Ant Colony Decision Tree (ACDT). Additionally, elements of Social Network Analysis (SNA) were included in this algorithm. A comparison of the results that have been obtained allows one to even more precisely determine the usefulness of the Ant Colony Optimization algorithm that is proposed here. It is one of artificial inteligence methods used in data mining [9]. The authors of the presented paper intend to show that emails can be classified into folders with a satisfactory accuracy, even if uncleaned data sets are included.

This article is organized as follows. Section 1 comprises an introduction to the subject of this article. In Sect. 2, characteristic of social networks is presented. Section 3 describes Ant Colony Decision Tree algorithm. Section 4 focuses on the presented, new version of the ACO approach based on ACDT algorithm and Social Network Analysis. Additionally, it shows the visualization of a network of contacts with objects. Section 5 presents the experimental study that has been conducted to evaluate the performance of the proposed algorithm, taking into consideration Enron e-mail dataset. Finally, we conclude with general remarks on this work and a few directions for future research are pointed out.

2 Characteristic of Social Networks

Social Networks Analysis (SNA) plays an extremely important role in studies of data sets containing email messages. Most of all, SNA provides a specific perspective on an analysis because it does not focus on individual units or macrostructures but studies the connections between particular units or groups.

SNA is a branch of sociology which deals with the quantitative assessment of the individuals role in a group or community by analyzing the network of connections between individuals. The first studies of social networks were conducted in 1923 by Jacob L. Moreno, who is regarded as one of the founders of social network analysis. Morenos 1934 book that is titled "Who Shall Survive?" presents the first graphical representations of social networks as well as definitions of key terms that are used in an analysis of social networks and sociometric networks [8].

A social network is usually represented as a graph. According to the mathematical definition, a graph is an ordered pair

$$G = (V, E), \tag{1}$$

where:
V denotes a finite set of a graphs vertices,
E denotes a finite set of all two-element subsets of set V that are called edges, which link particular vertices such that:

$$E \subseteq \left\{ \{u, v\} \, : \, u, v \in V, u \neq v \right\}. \tag{2}$$

Vertices represent objects in a graph whereas edges represent the relations between these objects. Depending on whether this relation is symmetrical, a graph which is used to describe a network can be directed or undirected.

The degree of a vertex (indegree and outdegree) denotes the number of head endpoints or tail endpoints adjacent to a given node. Degree centrality is useful in determining which nodes are critical as far as the dissemination of information or the influence exerted on immediate neighbors is concerned. Centrality is often a measure of these nodes popularity or influence.

Social network analysis has a wide range of applications. It is primarily used in large organizations and companies as a tool for supporting strategic human resource management or knowledge management in an organization. SNA supports a companys innovativeness and an analysis of business processes as well as training needs. Additionally, it is used in marketing research for creating a map of a social network of customers. However, social network analysis primarily allows managers to familiarize themselves with the informal structure of an organization and the flow of information within a company.

Many studies that were carried out as part of SNA were aimed at finding correlation between a networks social structure and efficiency [5]. At the beginning, social network analysis was conducted based on questionnaires that were filled out by hand by the participants [4]. However, research carried out by using email messages has become popular over time [1]. Some of the studies found that research teams were more creative if they had more social capital [6]. Social networks are also associated with discovering communication networks. The database which was used in the experiments that are presented in this article can be used to analyze this problem. G. C. Wilson and W. Banzhaf, among others, discussed such an approach, which they described in their article [10].

3 Ant Colony Decision Tree

The Ant Colony Decision Tree algorithm (ACDT) is one of the most popular Ant Colony Optimization algorithms that are used in data mining. This algorithm combines the idea of Ant Colony Optimization algorithms with the idea of the CART algorithm and, as tests have shown, it produces very good quality classifiers for many standard problems related to data mining [2]. The ACDT algorithm is based on using Ant Colony Optimization algorithms in the process of optimizing the construction of decision trees. The execution of the algorithm involves choosing a test for each node based on two factors. The maximum value that is consistent with the splitting criteria used in the CART algorithm is one of these factors, and the additional information that is recorded in the form of the pheromone trail is the other factor [2, 7].

In the Ant Colony System a virtual ant decides on the next step based on a modified transition rule while being at a particular stage of problem—solving and at a specific point in time. For this purpose, it generates a random number q, $0 \leq q \leq 1$. If $q \leq q_0$ (q_0— the parameter of the algorithm that has been determined) then "the best" available decision option is chosen (exploitation); otherwise, the ant makes a random decision (exploration) by taking into account probabilities that are calculated in accordance with Eq. (3) [7] (Table 1).

$$j = \begin{cases} \arg\max\limits_{r \in J_i^k} \{[\tau_{ir}(t)]^{\alpha} \cdot [\eta_{ir}]^{\beta}\}, & \text{if } q \leq q_0 \text{ (exploitation)} \\ S, & \text{otherwise (exploration)}, \end{cases} \tag{3}$$

where:
τ_{ir}—value of the reward, i.e. the degree of usefulness of the decision option that is being considered (pheromone),
η_{ir}—value of the quality of a transition from state i to state r which was estimated heuristically,
α i β—parameters describing the importance of values $\tau_{ir}(t)$ i η_{ir},
S—the next step (decision) which was randomly selected by using the probabilities:

$$p_{ij}^k(t) = \begin{cases} \frac{\tau_{ij}(t) \cdot [\eta_{ij}]^{\beta}}{\sum_{r \in J_i^k} \tau_{ir}(t) \cdot [\eta_{ir}]^{\beta}}, & \text{if } j \in J_i^k \\ 0, & \text{otherwise}, \end{cases}$$

where:
J_i^k denotes the set of decisions that ant k can make while being in state i.

The value of the heuristic function is determined based on the splitting criteria used in the CART algorithm, i.e. in accordance with the following Eq. (4).

$$\arg\max_{a_j \leq a_j^R, j=1,\ldots,M} \left(\frac{P_l P_r}{4} \left[\sum_{k=1}^{K} |p(k|m_l) - p(k|m_r)| \right]^2 \right), \tag{4}$$

where:
$p(k|m_l)$—probability of the occurrence of decision class k in node m_l (in the left subtree),
$p(k|m_r)$—probability of the occurrence of decision class k in node m_r (in the right subtree),

Table 1 Parameters in data sets

Dataset	N. of objects	N. of class	Number of attributes						Parameters of social networks	
			from	word1	word2	word3	cc	Length	N. of edges	Frequency of information flow
germany-c	1013	19	207	382	419	340	2	835	25	208
haedicke-m	112	16	64	70	85	67	2	110	30	238
mann-k	1616	24	254	394	490	433	2	1248	18	123
rogers-b	1395	14	289	445	521	430	2	1101	13	26
scott-s	641	10	135	350	219	166	2	578	28	443
shackleton-s	1001	53	158	330	384	357	2	836	27	1656
shapiro-r	1970	75	325	720	856	754	2	1566	25	1050
steffes-j	625	23	157	242	341	300	2	555	19	401
symes-k	770	12	119	324	346	287	2	685	16	387
taylor-m	656	21	173	255	288	245	1	580	32	1256

P_l—probability of object transition to node m_l (in the left subtree),
P_r—probability of object transition to node m_r (in the right subtree),
K—decision classes.
The pheromone trail is updated by increasing pheromone levels on the edges connecting each tree node with its parent node (excepting the root):

$$\Delta\tau_{m,m_L}(t+1) = (1-\gamma)\cdot\tau_{m,m_L}(t) + Q(T), \tag{5}$$

where:
$Q(T)$ determines the evaluation function of the decision tree,
γ is a parameter representing the evaporation rate, which is equal to 0.1.

4 Proposed Algorithm

The proposed method entails using a modified version of the ACDT algorithm (which is described in Sect. 3) and transforming a data set of emails into a decision table. For such a data set the proposed algorithm was prepared; it contains elements of communication network analysis which entails analyzing the list of recipients.

The decision table that has been prepared consists of the following attributes:

- from—the sender;
- word1—the first word which is used in the subject of an email (with the exception of basic words and copulas); additionally, words which belong to the set of decision classes are supported;
- word2—the second word which is established similarly to word1;
- word3—the third word which is established similarly to word1 and word2;
- cc—the Boolean value which indicates whether the person who has received an email was added as a recipient of a copy of an email (if not then it means that the person was the addressee of an email);
- length—number of characters of the mail (with white spaces);
- category—a decision class, i.e. a folder, to which an email message is assigned.

Conditional attributes were selected to define the most important information about each message. They consist of the information from the sender field, the first three words from the email subject, information as a Boolean value, conditional attributes check also the length of the message and whether the person who received the message was added to a courtesy copy (CC). If not, it implies that was the recipient (To). In addition, from the email subject was omitted basic phrases and copula verbs, and there was additionally supported words, which belonged to a set of decision classes.

Application of the ACDT algorithm (which is based on a modification of this algorithm at the present stage) entails exploring the communication network between people if an email was sent to a group of persons, i.e. $cc = true$. The list of all recipients is analyzed, which has an influence on which decision class (email folder) a classifier will choose.

This decision is also influenced by the preferences of the group of users who contact one another; therefore, if the users contacted one another with the same frequency then the emails they received were classified in the same way. Because of that the network of interactions has been analyzed using well-know methodology (from the SNA field).

All of the mailboxes that have been selected contain uncleaned data, which is why in the sets there might be folders without any messages or folders that were automatically created by email programs. Uncleaned data sets may also contain unnecessary email messages which were sent many times as well as emails that have not been assigned to any folder. The number of decision classes depends on the case, which is analyzed.

These data sets are very large - they are composed of a large number of decision classes and have attributes with many values, mainly with continuous values. Therefore, modified Ant Colony Decision Tree algorithm was used to analyze this data set because they perform very well as far as such problems are concerned [2]. The way in which such algorithm work is presented based on the example of Algorithm 1.

Algorithm 1: Pseudo code of the proposed algorithm

```
dataset = prepare_dataset_from_email(person);
ph = initialization_pheromone_trail(); // $\tau_{m,m_L}(t=0) = \frac{\log_2(C)}{\sum_{att=1}^{|A|} |a_{att}|}$
best_constr._classifier = NULL;
for i=1 to number_of_iterations do
    best_classifier = NULL;
    for j=1 to number_of_ants do
        new_classifier = build_prototype_classifier_EMAIL(ph, dataset);
        new_classifier = check_contacts_SNA(new_classifier, dataset);
        assessment_of_the_quality_classifier(new_classifier);
        if new_classifier is_higher_quality_than best_classifier then
            best_classifier = new_classifier;
        endIf
    endFor
    update_pheromone_trail(best_classifier, ph);
    if best_classifier is_higher_quality_than best_constr._classifier then
        best_constr._classifier = best_classifier;
    endIf
endFor
result = best_constr._classifier;
```

The proposed analysis of the network of contacts between individual employees is used to determine the leaders in terms of the spread of information or to influence the persons, who are in the immediate vicinity. Figure 1 shows a visualization of the network of contacts with objects selected from the dataset of 150 objects. The dataset was analyzed in terms of the frequency of sending e-mails. Additionally, Table 1 shows parameters for each dataset.

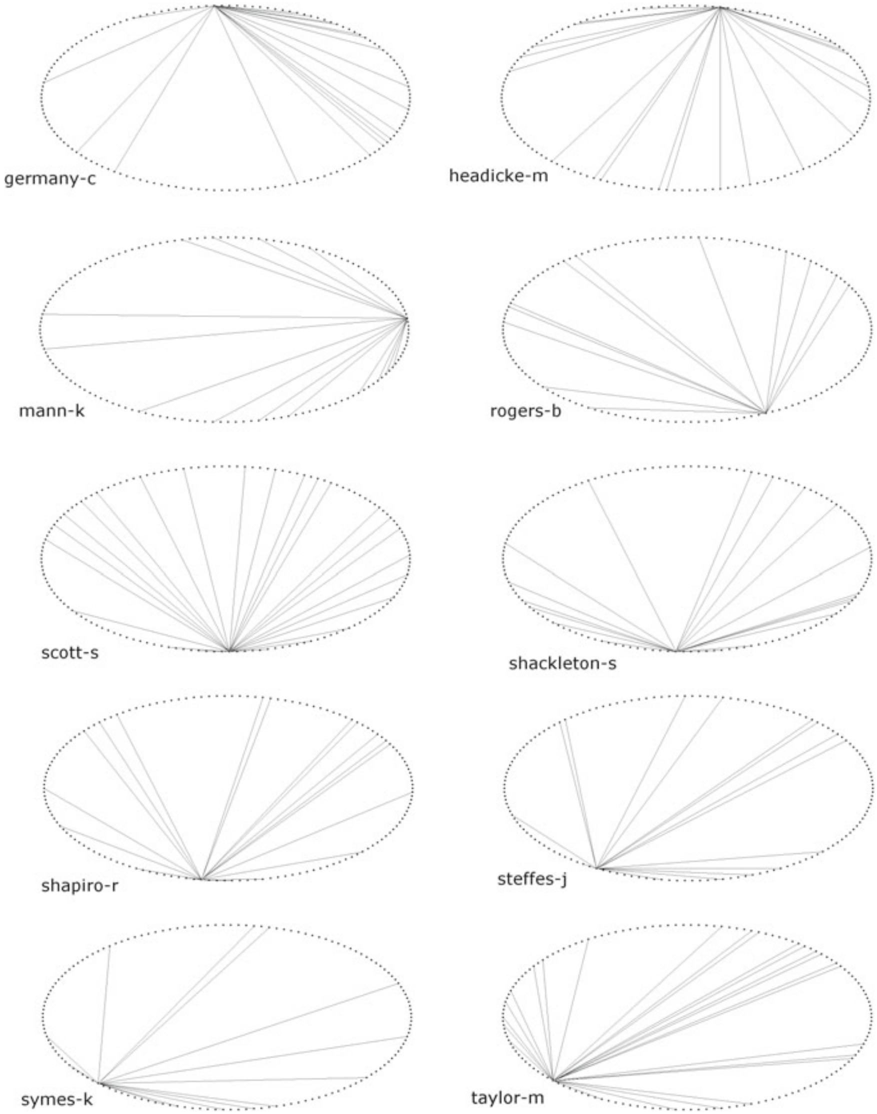

Fig. 1 Social networks for Enron dataset

5 Experiments

The proposed algorithm was implemented in C++. All computations were carried out on a computer with an Intel Core i5 2.5 GHz processor, 2.9 GB RAM, running on the Debian GNU/Linux operating system.

Table 2 Comparison of all approaches in terms of classification accuracy on uncleaned data sets

Dataset	Simple cart	Naive Bayes	Conjunctive rule	PART	Hyper pipes	LBR	Proposed algorithm
germany-c	0.386	0.626	0.415	0.576	0.549	0.626	**0.783**
haedicke-m	0.297	0.622	0.297	0.297	0.595	0.622	**0.623**
mann-k	0.245	0.708	0.178	–	0.684	0.712	**0.767**
rogers-b	0.510	0.772	0.443	0.746	0.738	0.774	**0.911**
scott-s	0.662	0.826	0.615	0.723	0.864	0.836	**0.936**
shackleton-s	0.565	0.667	0.291	0.682	0.628	0.673	**0.709**
shapiro-r	0.091	0.421	0.125	–	0.392	0.428	**0.605**
steffes-j	0.649	0.755	0.548	0.639	0.769	0.755	**0.841**
symes-k	0.324	0.789	0.457	0.723	0.781	0.785	**0.930**
taylor-m	0.367	0.757	0.399	0.321	0.757	0.761	**0.862**

The experiments were repeated 30 times for each data set with the same standard parameter settings which were related to Ant Colony Optimization algorithms. Given the size of the data set, the number of generations of the Ant Colony Optimization algorithm was initially restricted to 30 for a population of 5 ants. The run-time of the proposed algorithm ranged, depending on the data set, between 7 and 400 s for one run of the algorithm. This is, however, a time during which a classifier is created whereas classification itself is carried out very quickly.

In order to check the proposed algorithms adaptability, experiments were carried out so as to make it possible to compare the obtained results with the results that were produced by other classifiers. The algorithms that had been chosen constructed classifiers by using the same data sets as the SNA Ant Colony Optimization Algorithm that was proposed in [3]. Additionally, a larger number of data sets (email messages) were used to test the algorithms.

The experiments that are described in this section were conducted in order to check if the proposed method for classifying email messages into folders works correctly and also to verify whether the proposed method of creating decision tables can be used for any classifiers, even if uncleaned data sets are included. Out of all mailboxes obtained from the Enron E-mail data set a total of ten mailboxes were selected. The mailbox sizes ranged from 10 to 42MB, which could indicate that they contained a small number of emails and folders.

The proposed algorithm, which is described in Sect. 4, and selected algorithms that had been implemented in the Weka system (Waikato Environment for Knowledge Analysis) [11] were chosen for the purpose of carrying out the tests. All tests were conducted on uncleaned data sets. The obtained results are presented in Table 2 and Fig. 2.

The proposed algorithm each time generated better results for all data sets that had been created based on ten users. As for three data sets (mann-k, scott-s and steffes-j), the accuracy with which a folder is assigned to an email improved by 5–7 % points in

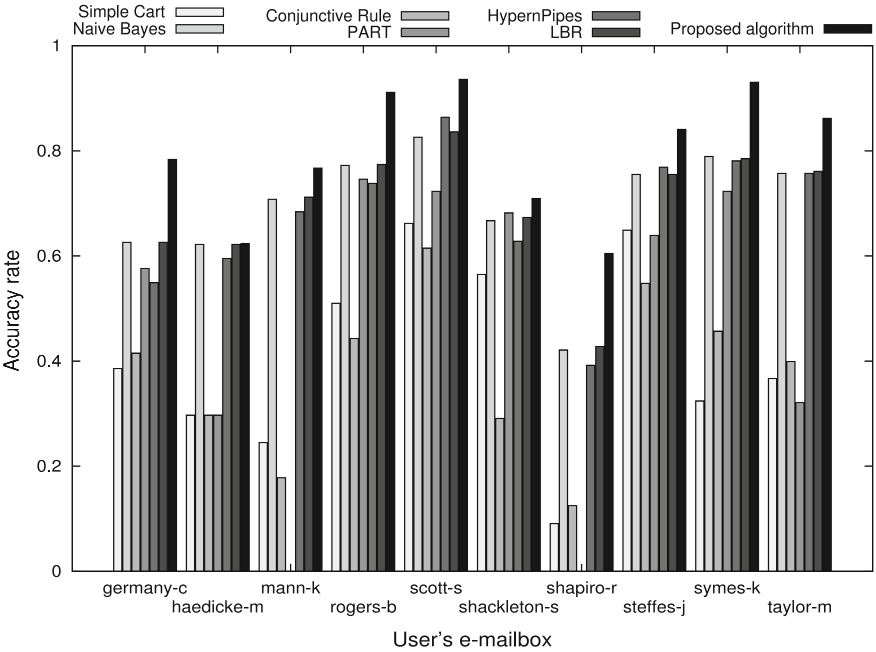

Fig. 2 The correctness of the proposed categorization method

relation to the best of the other methods that were compared. For three other data sets (rogers-b, symes-k and taylor-m) there was a large, 10–14 % points improvement, whereas for sets germany-c and shapiro-r there was a very large improvement, i.e. of 16–18 % points.

For the two remaining sets (haedicke-m and shackleton-s), all of the algorithms achieved the same level of classification accuracy. The results concerning classification accuracy achieved by using the PART algorithm are not provided for sets mann-k and shapiro-r because the run-time of this algorithm was too long.

Other elements that are related to analyzing algorithms also need to be compared, i.e. those which could not be compared at this stage. Nonetheless, the classification stage itself is very similar for all the methods; therefore, potential differences may only result from the complex structure of the classifiers.

6 Conclusions

Based on the experiments that were carried out, it was confirmed that the accuracy of classification, i.e. the correctness of an automatic categorization of email messages, was considerably improved when Ant Colony Optimization algorithms and Social Network Analysis were used. The aim of this article has been achieved.

The proposed approach led to a significant improvement in the classification of emails into folders. Creating a map of contacts in the form of social networks suggests that method will not only reduce the time spent on reading and replying to e-mails received, but above all is crucial to the process flow of information between employees of the company.

It has been noticed that the proposed method of creating decision tables makes it possible to use classical classifiers to categorize email messages. However, the proposed algorithm produces even better results due to its adaptability and the use of SNA elements.

In the future the authors of this article intend to adapt the social network mechanism for this purpose to a larger extent and to improve the process of creating decision tables. In future stages of the research, the incorporation of elements of text mining in an analysis of email message content and the direct coupling of these elements with the pheromone trail of the proposed algorithm should produce positive effects.

References

1. Aral, S., Van Alstyne, M.: Network structure & information advantage. In: Proceedings of the Academy of Management Conference, vol. 3, Philadelphia, PA. Citeseer (2007)
2. Boryczka, U., Kozak, J.: Ant Colony Decision Trees—a new method for constructing decision trees based on Ant Colony Optimization. In: Computational Collective Intelligence. Technologies and Applications, LNCS, vol. 6421, pp. 373–382. Springer (2010)
3. Boryczka, U., Probierz, B., Kozak, J.: An ant colony optimization algorithm for an automatic categorization of emails. Computational Collective Intelligence. Technologies and Applications, LNCS, vol. 8733, pp. 583–592. Springer, Berlin (2014)
4. Cummings, J.N., Cross, R.: Structural properties of work groups and their consequences for performance. Soc. Netw. **25**(3), 197–210 (2003)
5. Gloor, P.A.: Swarm creativity: competitive advantage through collaborative innovation networks. Oxford University Press (2005)
6. Gloor, P.A., Grippa, F., Putzke, J., Lassenius, C., Fuehres, H., Fischbach, K., Schoder, D.: Measuring social capital in creative teams through sociometric sensors. Int. J. Organ. Des. Eng. **2**(4), 380–401 (2012)
7. Kozak, J., Boryczka, U.: Enhancing the effectiveness of ant colony decision tree algorithms by co-learning. Appl. Soft Comput. **30**, 166–178 (2015)
8. Moreno, J.L.: Who shall survive? Foundations of Sociometry, Group Psychotherapy and Sociodrama. Beacon House (1953)
9. Tkacz, M.: Artificial neural networks in incomplete data sets processing. In: Intelligent Information Processing and Web Mining, pp. 577–583. Springer (2005)
10. Wilson, G., Banzhaf, W.: Discovery of email communication networks from the enron corpus with a genetic algorithm using social network analysis. In: IEEE Congress on Evolutionary Computation, 2009. CEC'09, pp. 3256–3263. IEEE (2009)
11. Witten, I.H., Frank, E., Hall, M.A.: Data Mining: Practical Machine Learning Tools and Techniques, 3rd edn. Morgan Kaufmann Publishers Inc. (2011)