ALGORITHMS FOR EFFECTIVE TEAM BUILDING

Sashka Ivanovska
Faculty of Electrical Engineering and Information Technologies
Skopje, R. Macedonia

Ilinka Ivanoska
Faculty of Computer Science and Engineering
Skopje, R. Macedonia

Slobodan Kalajdziski
Faculty of Computer Science and Engineering
Skopje, R. Macedonia

Abstract

Effective team building is an important issue of human resource management (HRM). In order to keep up with technological improvements and changes, selecting the right person for the right job position is very important. This paper describes a research and development methodology for establishing a more sophisticated approach for composing effective teams.

Data mining (DM) techniques and algorithms, like decision trees, Bayesian networks and fuzzy logic, were utilized to build a model to predict the best possible person for a specific job. We have applied K-means and fuzzy C-means clustering and decision tree classification algorithms. Pruned and unpruned trees were contributed using ID3, C4.5 and CART algorithms. By using these techniques, the patterns of employee performance were generated. To validate the generated model, several experiments were conducted using data collected from IT companies. After evaluation, the most appropriate algorithms are recommended to be used in the process of effective team building.

I. Introduction

In the recent years the significant changes are done in every industry which has important implications on data mining (DM). Organization and companies are using information technology (IT) for generating, storing and analyzing mass produced data, not only for operational purposes, but also for enabling strategic decision making to survive in a competitive and dynamic environment. DM helps in reducing information overload along with the improved decision-making by searching for relationships and patterns from the huge dataset collected by organizations. It enables company to focus on the most important information in the database and allows retailers to make more knowledgeable decisions by predicting future trends and behaviours.

In addition, today's organization has to struggle effectively in terms of cost, quality, service or innovation. The success of these tasks depends on having enough right people with the right skills, deployed in the appropriate locations at the appropriate point [1]. Well-trained and dedicated workers directly affect the success of the company, so employers introduce different techniques when selecting employees. Particular attention is paid to the selection of employees in building teams. Teams are an integral part of almost every company. The team is actually a group of people who share the same responsibility for a given activity and whose interaction and cooperation affect the successful implementation of the activity. Building effective teams is a complex process, as it requires a large amount of data manipulation: personal and technical characteristics of the human resources in companies; customer characteristics; characteristics of the projects. At the beginning, the process of building teams has been performed manually by the HRM experts. Manual selection of employees is impractical because it takes a lot of resources and a profound impact on the choice has the subjectivity of the person who made the choice.

The implementation of information technology in the fields of HRM, the process of building an effective team can automate to reduce costs and improve quality [2]. Techniques of DM, such as Bayes' theorem, clustering, association rules, prediction and classification, are best for analyzing data and generating schemes that can be used in the future. These techniques are commonly used for efficient and effective decision making in the process of building effective teams.

Recently, there are some researches that show great interest on solving HR problems using data mining approach [3]. In addition, until now there are quite limited discussions on human resources problems such as for talent forecasting, employee development [4, 5] and performance evaluation [6] use data mining approach. In HR, data mining technique used focuses on personnel selection [7, 8] especially to choose the right candidates for a job. The classification and prediction in data mining for HR problems are infrequent and there are some examples such as to predict the length of service, sales premium, to persistence indices of insurance agents and analyze miss-operation behaviors of operators [9]. Due to these reasons, this study attempts to use data mining classification techniques to effective team building using past experience knowledge.

The aim of this paper is to evaluate the DM algorithms and techniques which can be used for effective team building. In order to reach the goal, tests have been made on the dataset that is a representative of a dataset of an existing organization. According to the results from the testing a comparison between the algorithms is made, and their advantages and disadvantages are described.

As a final result of this paper, the most appropriate algorithms are recommended to be used in the process of effective team building.

II. Overview of algorithms for Effective Team Building

A. K-means Clustering

Unsupervised K-means learning algorithms are used for solving the well known clustering problem. The algorithm aims at forming *k* clusters of *n* objects such that the resulting intra-cluster similarity is high but the inter-cluster similarity is low. The algorithm randomly selects *k* of the *n* objects and one of them is assigned to each cluster to represent the cluster mean or the center. For each of the remaining objects, an object is assigned to the cluster to which it is most similar, based on the distance between the object and the cluster mean. Then new mean is computed for each cluster and the process iterates until the criterion function converges. A square-error criterion is used and defined as (1) [10].

$$E = \sum_{i=1}^{k} \sum_{p=Ci} | p - m_i |^2 \tag{1}$$

Place *k* points into the space represented by the objects that are being clustered. These points represent initial group centroids.

Assign each object to the group that has the closest centroid.

When all objects have been assigned, recalculate the positions of the *k* centroids.

Repeat Steps 2 and 3 until the centroids no longer move. This produces a separation of the objects into groups from which the metric to be minimized can be calculated.

B. ID3 Algorithm

ID3 is a simple decision learning algorithm that uses statistical property call information gain to select which attribute to test at each node in the tree. Information gain measures how well a given attribute separates the training examples according to their target classification [11].

Entropy is a measure in the information theory, which characterizes the impurity of an arbitrary collection of examples. If the target attribute takes on c different values, then the entropy S relative to this c-wise classification is defined as (2)

$$Entropy(S) = \sum_{i=1}^{c} - p_i \log_2 p_i \tag{2}$$

where p_i is the proportion/probability of *S* belonging to class *i*. Logarithm is base 2 because entropy is a measure of the expected encoding length measured in bits.

The information gain, *Gain(S, A)* of an attribute *A*, relative to the collection of examples *S*, is defined as (3)

$$Gain(S, A) = Entropy(S) - \sum_{v \in Value(A)} \frac{|S_v}{|S|} Entropy(S_v) \tag{3}$$

where *Values*(*A*) is the set of all possible values for attribute A, and S_v is the subset of *S* for which the attribute *A* has value *v*. We can use this measure to rank attributes and build the decision tree where at each node is located the attribute with the highest information gain among the attributes not yet considered in the path from the root.

C. C4.5 Algorithm

C4.5 algorithm is a successor of ID3 that uses gain ratio as splitting criterion to partition the data set. The algorithm applies a kind of normalization to information gain using a "split information" value (4) [12].

$$Split(T) = -\sum_{i=1}^{s} \frac{|T_i|}{|T|} x \log_2 (\frac{|T_i|}{|T|}) \tag{4}$$

where |*Ti*| is the number of instances in the training set *T* with it value for the attribute *T* and |*T*| is the total number of instances in the training set. Gain ratio is defined as in (5) and the attribute with maximum gain ratio is selected as the splitting attribute.

$$\text{Gainratio}(T) = \frac{Gain(T)}{Split(T)} \tag{5}$$

D. CART Algorithm

CART is a recursive partitioning method that builds classification and regression trees for predicting continuous dependent variables and categorical predictor variables. The fundamental idea is to select each split of a subset so that the data in each of the descendant subsets are purer than the data in the parent subset. Main Steps for making a decision tree using CART Algorithm [13]:

The first is how the splitting attribute is selected.

The second is deciding upon what stopping rule need to be in place.

The last is how nodes are assigned to classes.

E. Bayesian Network

A Bayesian network encodes the joint probability distribution of a set of *v* variables, as a directed acyclic graph and a set of conditional probability tables.

Bayes' theorem addressed both the case of discrete probability distributions of data and the more complicated case of continuous probability distributions. In the discrete case, Bayes' theorem relates the conditional and marginal probabilities of events A and B, provided that the probability of B does not equal zero (6) [14]:

$$\mathrm{P}(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)} \tag{6}$$

In Bayes' theorem, each probability has a conventional name:

- *P(A)* is the prior probability (or "unconditional" or "marginal" probability) of A. It is "prior" in the sense that it does not take into account any information about B; however, the event B need not occur after event A.
- *P(A|B)* is the conditional probability of A, given B. It is also called the posterior probability because it is derived from or depends upon the specified value of B.
- *P(B|A)* is the conditional probability of B given A. It is also called the likelihood.
- *P(B)* is the prior or marginal probability of B, and acts as a normalizing constant. Bayes theorem in this form gives a mathematical representation of how the conditional probability of event A given B is related to the converse conditional probability of B given A.

The initial development of Bayesian networks in the late 1970s was motivated by the need to model the top-down (semantic) and bottom-up (perceptual) combination of evidence in reading. The capability for bidirectional inferences, combined with a rigorous probabilistic foundation, led to the rapid emergence of Bayesian networks as the method of choice for uncertain reasoning in Artificial Intelligence (AI) and expert systems replacing earlier, ad hoc rule-based schemes.

III. Problem Description

Teams are an integral part of almost every organization. The use of teams allows consideration of expertise in multiple areas as team members are often brought together with diverse knowledge, skills, and abilities.

The first and most important step in building teams is a detailed analysis of the project and preparation of specific technical specification for the client's demands [15]. In order to bring a project to realization, a construction of team made of top experts in their field is needed. For that purpose, employees' profiles that will be part of the team to be engaged in the realization of the project are being defined.

The profile defines the technical, social and organizational characteristics that employees to be part of the team should have. Several methods that give satisfactory results for automatic employee selection are being developed and are being used in most of the existing systems [16, 17, 18, 19]. In the last few years, the work on the improvement of the current techniques is being made, in order to gain better results that will directly affect the final selection.The fuzzy data model [20] is the most used model in employee selection. With this method, new employee pre-selection for already defined working profiles is being done. The model analyses the attributes of the employees and compares them with the attributes from the defined profiles. In this way, a fair and independent applicant selection is ensured; who will be later further interviewed in order to assess their additional characteristics.

Grouping the employees using the clustering and classifying techniques is another phase in the building of an effective team. Clustering is a process of segregation of the data set in several sub groups on the basis of some similar characteristics. Clustering enables the use of several attributes in the identification of similar characteristics in a group. The biggest problem with this technique is the selection of attributes because the performances of the algorithm are proportional to the number of attributes [21].

Also, the fuzzy C-means and K-means clustering algorithms have been used [22] in effective team building. The main idea implemented in these algorithms is a non-uniform division of data in a collection of clusters, which increases the freedom in the selection of most appropriate cluster.

In the process of creation of classification rules, usually the decision trees that are most simple for interpretation and understanding are being used [23]. The most famous algorithms that use the logic of decision trees, ID3, C4.5 and CART are being successfully used in the process of effective team building [24, 25]. These algorithms obtain accurate and precise results from the data analysis, but still, the extent of successful application of these algorithms depends on the data structure.

The division of employees in teams can be also made with the use of classification techniques like Bayes nets, associative rules and neural networks. Bayes nets are used for classification of structured data, and usually use stochastic algorithms [26]. Associative rules on the other hand are used in the process of data reduction, which significantly facilitates the further process of classification [27]. The benefit of using neural nets is that their application does not require prior training of the data set [28].

For a team to function efficiently, except for the process of team formation, the process of modifying its structure is of utmost importance as well. If a team member is off, it is necessary to fill his place with another employee whose expert characteristics are in accordance with the working position. The conducted researches in this area use hybrid models of multiple algorithms for more successful project realization [29, 30].

IV. Data Sets

When the idea of the study came into mind, it was intended to apply a classification model for predicting performance depending on a dataset from a certain IT company. So that any other factors regarding the working environment, conditions, management and colleagues would have similar effect on all employees, and so the effect of collected attributes would be more apparent and easier to classify. Unfortunately, data collected from the real IT Company was not enough to be the base of such a classification model. In this case, another attempt was taken to collect another group of data from another IT company. In order to collect the required data, we decided to use the dataset from data mining challenge [31]. Dataset contains one table with details of the employees and one table that contain details of the project requirements. The first dataset includes the historical employees details and consists of 2800 records, the second dataset includes the details of the project task and consists of 3500 records. From the dataset, it is observed, that the dataset consists of more than 95% of records to be in the rejected category. Hence the machine learning algorithms were very excellent in recognizing the rejected data however they were not able to identify selected records to a large extent. Therefore the dataset was premeditated, and almost 758 records in both categories were used for experimentation.

Figure 1 shows the process of obtaining results from this dataset which includes: data preparation, model creation and evaluation of results.

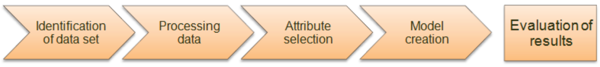

Figure 1: Data preparation, model creation and evaluation of results

When the records were chosen for the learning process, the distribution of the status in the original data was maintained.

The dataset tables were prepared and converted to (csv) format to be compatible with the WEKA data mining toolkit, which is used in the model building [32]. The algorithms were trained and tested with the 10-folds cross validation technique. The 10-folds cross validation is implemented to acquire the correct percent. For each 10-folds cross validation, the data set was first partitioned into 10 equal sized sets and each set was then in turned used as the test set while the system acquires rules from the other nine sets [33].

V. Experiments and Results

After the data has been prepared, the classification models have been built. Five classification techniques have been applied on the dataset on hand to build the classification model. The techniques are: The decision tree with two versions, ID3 and C4.5 (J4.8 in WEKA), CART, Fuzzy K-means and Bayes classifier. For each experiment, ROC Area [34] was evaluated using 10-folds cross validation. The results are presented in Table 1 and Table 2.

Table 1: The results using full attributes

Algorithm	ROC Area Performance
Bayes classifier	0.925
C4.5	0.847
ID3	0.782
Fuzzy K-means	0.630
CART	0.599

Table 2: The results using selected attributes

Algorithm	ROC Area Performance
Bayes classifier	0.926
C4.5	0.847
ID3	0.798
Fuzzy K-means	0.662
CART	0.599

It was observed that the ROC Area performance of Bayes classifier was better than other classification techniques. The constructed trees were used to study the impact of the input attributes. It may be observed from Table 1 and Table 2 that ID3 algorithm has a poor performance and the tree constructed with the C4.5 algorithm has better ROC Area performance. Generated rules were explored to determine the attributes that impact the team building process.

VI. Conclusion & Future work

In this paper, we have described the significance of the study on the use of data mining algorithms for effective teams building. Traditional methods of assembling teams are based on manual matching of individual skills to the requirements of the project tasks. These methods often do not consider the quality and mix of the teamwork and collaboration skills across the team members. By automating the building of teams and considering taskwork and teamwork competencies, the team members will be better prepared tackle the mission goals in a collaborative manner.

We proposed sophisticated team composition algorithms which represent a more robust method to index, study, and predict team performance. The underlying datasets have been analyzed to identify the best algorithms for creating the model. Most popular clustering and classification techniques were deployed in solving the problem. It was observed that Fuzzy C-means and K-means clustering techniques are not suitable for this type of data distribution. The three popular decision tree construction algorithms, ID3, C4.5 and CART have been applied. It has been observed that rule generated with Bayesian Network has better ROC Area performance. Analysis has been made on the constructed tree to deduce viable rules.

The future directions of the research is to investigate other DM algorithms for analyzing various types of data. As future work, it is recommended to collect more proper data from several companies. Databases for current employees and even previous ones can be used, to have a correct performance rate for each one of them.

REFERENCES

[1] H. Jantan, A.R. Hamdan, and Z.A. Othman, “Knowledge Discovery Techniques for Talent Forecasting in Human Resource Application”, *International Journal of Human and Social Sciences*, 5:11, 2010

[2] M J. Kavanagh and M. Thite, “Human Resource Information Systems”, p.3-22, 2009

[3] J. Ranjan, “Data Mining Techniques for better decisions in Human Resource Management Systems”, *International Journal of Business Information Systems*, vol. 3, no. 5, pp. 464-481, 2008

[4] K.Y. Tung, I.C. Huang, S.L. Chen and C.T. Shih, “Mining the Generation Xer's job attitudes by artificial neural network and decision tree - empirical evidence in Taiwan”, *Expert Systems and Applications*, vol. 29, no. 4, pp.783-794, 2005

[5] M. J. Huang, Y.L. Tsou and S.C. Lee, “Integrating fuzzy data mining and fuzzy artificial neural networks for discovering implicit knowledge”, *Knowledge-Based Systems*, vol. 19, no. 6, 396-403, 2006

[6] C. Xiaofan, “Application of Data Mining on Enterprise Human Resource Performance Management”, *International Conference on Information Management, Innovation Management and Industrial Engineering (ICIII)*, vol. 2, pp. 151 – 153, 2010

[7] M. Dursun, E.E. Karsak, “A fuzzy MCDM approach for personnel selection”, *Expert Systems with Applications*, vol. 37, pp. 4324–4330, 2010

[8] Q.A. Al-Radaideh, E. Al Nagi, “Using Data Mining Techniques to Build a Classification Model for Predicting Employees Performance”, *International Journal of Advanced Computer Science and Applications*, vol. 3, no. 2, 2012

[9] C.F. Chien and L.F. Chen, “Data mining to improve personnel selection and enhance human capital: A case study in high technology industry”. *Expert Systems and Applications,* vol. 34, no. 1, pp. 380-290, 2008

[10] Ch.G.V.N. Prasad, K.H. Rao, D. Pratima and B.N. Alekhya, “Unsupervised Learning Algorithms to Identify the Dense Cluster in Large Datasets”, *International Journal of Computer Science and Telecommunications*, vol. 2, no. 4, July 2011

[11] A. Bahety, “Extension and Evaluation of ID3 – Decision Tree Algorithm”, Department of Computer Science University of Maryland, College Park [Accessed February 2013]

[12] M.M. Mazid, S. Ali and K.S. Tickle, “Improved C4.5 Algorithm for Rule Based Classification”, School of Computing Science, Central Queensland University, Australia

[13] S. Soni, “Implementation of Multivariate Data Set by CART Algorithm”, *International Journal of Information Technology and Knowledge Management*, vol 2, no. 2, pp. 455-459, 2010

[14] D. Grossman and P. Domingos, “Learning Bayesian Network Classifiers by Maximizing Conditional Likelihood”, Department of Computer Science and Engineeering, University of Washington, Seattle [Accessed February 2013]

[15] C.W. Fontaine, “Working in teams-The Basics”, Prentice Hall, 2007

[16] T. Dereli, A. Durmuşoğlu, S.U. Seçkiner and N. Avlanmaz, “A fuzzy approach for personnel selection process”, *An Official Journal of Turkish Fuzzy Systems Association*, vol.1, no.2, pp. 126-140, 2010

[17] M. Othman, K.R. Ku-Mahamud and A.A. Bakar, “Fuzzy evaluation method using fuzzy rule approach in multicriteria analysis”, *Yugoslav Journal of Operations Research*, 2008

[18] H.T. Lin, “Personnel selection using analytic network process and fuzzy data envelopment analysis approaches”, Elsevier Ltd, 2010

[19] U. Straccia, E. Tinelli, S. Colucci, T.Di Noia and E.Di Sciascio, “A System for Retrieving Top-k Candidates to Job Positions”, *ISTI-CNR*, 2009

[20] W. Tai and C. Hsu, “A Realistic Personnel Selection Tool Based on Fuzzy Data Mining Method”, *National Yunlin University of Science and Technology*, 2006

[21] D. Perera, J. Kay, I. Koprinska, K. Yacef and O.R. Zaiane, “Clustering and Sequential Pattern Mining of Online Collaborative Learning Data”, *Knowledge and Data Engineering IEEE Transactions on*, 21(6), pp. 759-772, 2004

[22] O. Kaya, “ Human Resource Performance Clustering by Using Self. Regulating Clustering Method”, Master thesis, 2008

[23] J. Ranjan and R. Agarwal, “Advantages Of Decision Trees Using Data Mining In Indian Retail Industry”, *Journal of Knowledge Management Practice*, vol. 11, Special Issue 1, 2010

[24] N. Sivaram and K. Ramar, “Applicability of Clustering and Classification Algorithms for Recruitment Data Mining”, *International Journal of Computer Applications*, vol. 4, no.5, 2010

[25] H. Jantan, A.R. Hamdan and Z.A. Othman, “Classification Techniques for Talent Forecasting in Human Resource Management”, *InTech*, 2011

[26] E. Gyftodimos and P.A. Flach, “Learning Hierarchical Bayesian Networks for human skill modelling”, Department of Computer Science, University of Bristol, 2005

[27] J. Han and Y. Dong, “Application of Association Rules Based on Rough Set in Human Resource Management*”, The Sixth Wuhan International Conference on e-business* (WHICEB2007), 2007

[28] L.C. Huang, P. Wu, R.J. Kuo and H.C. Huang, “A neural network modeling on human resource talent selection”, *International Journal of Human Resources Development and Management*, vol.1, no.2/3/4, 2001

[29] N. Sivaram and K. Ramar, “Applicability of Clustering and Classification Algorithms for Recruitment Data Mining*”, International Journal of Computer Applications*, vol.4, no.5, 2010

[30] S.I. Tannenbaum, J.S. Donsbach, and G.M. Alliger, “Forming effective teams: testing the team composition system (TCS) algorithms and decision aid”, US Army Research Institute, W91WAW-08-C-0021

[31] https://www.kaggle.com/c/job-recommendation

[32] http://www.cs.waikato.ac.nz/ml/weka

[33] C. Carlsson and R. Fuller, “Fuzzy reasoning in decision making and optimization”, New York: Physica-Verl., 2002

[34] M. Zhu, “Recall, Precision and Average Precision”, Department of Statistics and Actuarial, Science University of Waterloo, 2004