Expert Systems with Applications 42 (2015) 1436–1445

Contents lists available at ScienceDirect

Expert Systems with Applications

journal homepage: www.elsevier.com/locate/eswa

From data mining to knowledge mining: Application to intelligent agents

Amine Chemchem *, Habiba Drias *

USTHB-LRIA, BP 32 El Alia Bab Ezzouar, Algiers, Algeria

ARTICLE INFO

Article history:
Available online 10 September 2014

Keywords:
Knowledge mining
Induction rules
Classification
Clustering
Cognitive agent

ABSTRACT

The last decade, the computers world became a huge wave of data. Data mining tasks were invoked to tackle this problem in order to extract the interesting knowledge. The recent emergence of some data mining techniques provide also many interesting induction rules. So, it is judicious now to process these induction rules in order to extract some new strong patterns called meta-rules. This work explores this concept by proposing a new support for induction rules clustering and classification. The approach invokes k-means and k-nn algorithms to mine induction rules using new designed similarity measures and gravity center computation. The developed module have been implemented in the core of the cognitive agent, in order to speed up its reasoning. This new architecture called the Miner Intelligent Agent (MIA) is tested and evaluated on four public benchmarks that contain 25,000 rules, and finally it is compared to the classical one. As foreseeable, the MIA outperforms clearly the classical cognitive agent performances.

1. Introduction

Nowadays, the induction rules have become inseparable pattern of the artificial intelligence thanks to their existence as the basis for many disciplines, such as the agent technology, data mining and knowledge discovery...This paper is about how to extend data mining techniques to induction rules in order to extract meta-rules. There are many data mining tasks for instance: clustering, classification, association rules mining, regression, prediction,...We are interested through this work in the first two tasks which are used on many applications (the image processing, the intrusion detection,...etc) and can be solved by different algorithms (k-means, HCA, fuzzy c-means...for clustering, KNN, SVM,ID3,...for classification). K-nn and K-means are in the top ten of data mining algorithms (Wu et al., 2008). The latter are extended to induction rules by introducing new version of similarity measure and gravity center computation. The algorithms called K-NN-IR and K-means-IR are developed and demonstrated on a public large scale benchmark including 25,000 induction rules. The whole idea behind this work is to improve the reasoning process by integrating the knowledge mining module in today's intelligent agent in order to speed up the reasoning engine process.

The rest of this paper is organized as follows: Next section shows a short history of data mining. Section 2 summarizes related works. In Section 3: induction rules representation are presented, followed by proposing mathematical preliminaries. In Section 5, the suggested algorithms are described and followed by the definition of a new architecture for intelligent agent. Then, experimental results are shown in Section 7 compared to the previously proposed algorithms. Finally we conclude by making some remarks and talking about future works.

2. Data mining overview

The generation of models from a large number of data is not a recent phenomenon. Egypt Pharaoh Amasis organizing the census of the population in the fifth century BC Rocchi (Rocchi, 2003). This is the seventeenth century we begin to analyze the data to find common characteristics. In 1662, John Graunt published his book "Natural and Political Observations Made upon the Bills of Mortality" in which he analyzed the mortality in London and trying to predict the appearances of the bubonic plague. In 1763, Thomas Bayes shows that we can determinate not only probabilities from observations derived from experience, but also the parameters for these probabilities. Legendre published in 1805 an essay on the least squares method for comparing a set of data with a mathematical model. From 1919 to 1925, Ronald Fisher develops the analysis of variance as a tool for its proposed medical statistical

* Corresponding authors.
E-mail addresses: aminechemchem@gmail.com (A. Chemchem), hdrias@usthb.dz (H. Drias).

http://dx.doi.org/10.1016/j.eswa.2014.08.024
0957-4174/

inference. The 1950s saw the advent of computer technology and computer calculation. Same methods and techniques are emerging such as segmentation, neural networks and genetic algorithms, and then in the 1960s, the decision tree, the method of mobile centers, these techniques allow researchers to exploit and discover models more accurate. The advent of the microcomputer stimulates research and statistical analyzes are more numerous and precise. The term "data mining" had a negative connotation in the early 1960s, expressing contempt for statisticians research approaches without correlation assumptions. it fell into oblivion, and Rakesh Agrawal employed again in the 80s when they were beginning research on databases with a volume of 1 Mb. The concept of data mining makes its appearance – according Pal (2007) – when the IJCAI[1] conferences took place in 1989. Then, in the 1990s, came the machine learning techniques such as SVM in 1998, complementing the tools of the data analysis. At the turn of the century, a company like Amazon uses these tools to offer our customers products that may interest. Actually, There are many tasks of data mining such as: Supervised and unsupervised classification, association rule mining, prediction and regression.

2.1. Supervised classification

Classification of a collection consists of dividing the items that make up the collection into categories or classes (Kotsiantis, 2007; Jain, Murty, & Flynn, 1999). In the context of data mining, classification is done using a model that is built on historical data. The goal of predictive classification is to accurately predict the target class for each record in new data, that is, data that is not in the historical data. A classification task begins with build data (also known as training data) for which the target values (or class assignments) are known. Different classification algorithms use different techniques for finding relations between the predictor attribute's values and the target attribute's values in the build data. K Nearest Neighbor (K-NN from short) is one of those algorithms that are very simple to understand, furthermore, it works incredibly well in practice, especially in the anomaly detection domain like Liao and Vemuri (2002), also for text categorization like in the work Guo, Wang, Bell, Bi, and Greer (2006). Also it is surprisingly versatile and its applications range from vision to proteins to computational geometry to graphs and so on. With KNN algorithm, we can obtain a satisfactory results, in addition, its basic principle is very simple, and easy to implement. It also might surprise many to know that K-NN is one of the top 10 data mining algorithms. K-NN is an non parametric learning algorithm, it is used when the data set does not obey a defined function as (gaussian mixtures, linearly separable etc). K-NN algorithm can be explained as follows, in the first time, training data that are already classified are considered, and then to classify the new data, we have to compute the similarities distance between this new data and all training data. After that the k nearest neighbors are extracted. In the end the new data is assigned to the most frequent class of these neighbors.

2.2. Clustering data technique

Clustering data mechanism consist to put the homogeneous data into the same group or class in order to dispatch the heterogeneous data into different groups. In the literature, it exists different manner to group the data, the two principals are: the hierarchical and the partitioning clustering. For the hierarchical clustering, the clusters are inside each others. This category of clustering is used when data can be separated in different levels. Also, CHA is the most known hierarchical algorithm, it starts by putting each instance in one cluster after that it computes the dissimilarities for all two instances to combine the clusters that have the lower distance. This process is repeated until we get one cluster (Steinbach, Ertöz, & Kumar, 2004; Han, Kamber, & Pei, 2006). In the contrary of the partitioning clustering, it consists to cluster the data separately. K-means is one of the simplest pure partitioning learning algorithms that solves the well known clustering problem (Han et al., 2006; MacQueen et al., 1967). The procedure follows a simple and easy way to classify a given data set through a certain number of clusters (assume k clusters) fixed initially. The main idea is to define k gravity centers, one for each cluster. The centroids should be placed in a cunning way because the clustering result depends on their location in the clusters. In order to optimize the efficacy of the outcomes, it is judicious to place them as much as possible far away from each other. The next step is to take each point belonging to a given data set and associate it to the nearest centroid. When no point is pending, the first step is completed and an early grouping is done. At this point we need to recalculate k news centroids of the clusters resulting from the previous step, and iterates the process. The latter stops when no more changes of the clusters are observed, in other words when the centroids do not move any more.

3. Related works

Our interest in this study revolves around two main subjects: scalable cognitive agent and knowledge mining in general which involves induction rules mining. As for first subject, we found very few papers with ideas about the notion of scalable cognitive agent like Cao, Gorodetsky, and Mitkas (2009), and nothing about the paradigm that we would like to cover in this article. However, we notice that biologists and psychologists are showing interest in the study of scalable brain (Eliasmith, 2013). What can be said about the second topic is that the literature offers a large spectrum of detailed research on knowledge Mining. Mining knowledge including simple data and other patterns have been examined intensively over the last decade. In the following, we will talk about some knowledge mining.

Many works are about mining association rules in order to obtain meta rules whose purpose is to reduce the large number of discovered rules. The CLOSET algorithm was proposed in Strehl, Gupta, and Ghosh (1999) as a new efficient method for mining closed itemsets. CLOSET uses a novel frequent pattern tree (FP-tree) structure, which is a compressed representation of all the transactions in the database. Moreover, it uses a recursive divide-and-conquer and database projection approach to mine long patterns. Another solution for the reduction of the number is introduced by Hahsler and Chelluboina (2011) used an itemset-tid set search tree and pursued with the aim of generating a small non redundant rule set. To this goal, the authors first found minimal generator for closed itemsets, and then, they generated non redundant association rules using two closed itemsets. A new algorithm to group rules via hierarchical clustering has been developed in Berrado and Runger (2007) to visualize the large number of rules. The clustering of rules is done by defining a new distance called $d_{Jaccard}$ that represents the number of items of the two rules divided by the number of unique items. Saneifar, Bringay, Laurent, and Teisseire (2008) were interested in discovering sets of data. In their paper, they have developed a new similarity measure between two rules and extended k-means algorithm to cluster them. In literature some works about induction rules analysis have been proposed: In Poongothai and Sathiyabama (2012b), eh authors have developed a new algorithm to select the interesting induction rules from all the discovered rules in web mining

[1] International Joint Conference on Artificial Intelligence.