P: ISSN No. 2231-0045 RNI No. UPBIL/2012/55438 VOL.-III, ISSUE-III, February-2015
E: ISSN No. 2349-9443 Periodic Research

Algorithms in Data Mining: A Survey

Abstract

Data mining is the process of automatically finding useful information in large data repositories. The purpose of deploying data mining algorithms is discovering important patterns from database and also provides capabilities to calculate the outcome of a future study.

In this paper we discussed different data mining algorithms C4.5, k-Means, SVM, AdaBoost. These algorithms the most dominant data mining algorithms in the research area. With each algorithm, we provide a description of the algorithm discuss the impact of the algorithm, and Future research on the algorithm. These algorithms include classification, clustering, statistical learning, association analysis, and link mining, which are all among the most important area in data mining research and development.

Keywords: C4.5, Data Mining, SVM, K-Means, AdaBoost.

Archana Chhikara
Assistant Professor,
Deptt. of Computer Science,
Hindu College, Sonepat

Amita Gandhi
Assistant Professor,
Deptt. of Computer Science,
Hindu College, Sonepat

Introduction

We required a proper mechanism for collecting data from different large repositories for good decision making. Knowledge discovery in databases (KDD), often called data mining. The aim of data mining is discovery of useful information from large collections of data. Data mining is the process of analyzing data from different view and summarizing it into useful information, that is used for good decision making. Data mining software is one of a number of analytical tools for analyzing data. It allows users to analyze data from many different dimensions or angles, categorize it, and summarize the relationships identified. Technically, data mining is the process of finding patterns from different fields in large relational databases. There are many data mining tools available in market. In this Paper we described four most demanding data mining tools.

This paper is organized as follows section 1 presents C4.5algorithm section 2 prgesent k-Means algorithms section 3 present SVM algorithm section 4 present AdaBoost algorithm and at last conclusion.

C4.5 (Introduction)

C4.5 algorithm use concept of classification. Concept of classification one of the commonly used tool in data mining. Classification is a data mining technique that maps data into predefined groups or classes. It is a supervised learning method which requires labelled training data to generate rules for classifying test data into predetermined groups or classes. It is a two-phase process. The first phase is the learning phase, where the training data is analyzed and classification rules are generated. The next phase is the classification, where test data is classified into classes according to the generated rules. Such systems take as input a collection of cases, each belonging to one of a small number of classes and described by its values for a fixed set of attributes, and output a classifier that can accurately predict the class to which a new case belongs.C4.5 generates classifiers expressed as decision trees, it also construct classifiers in more comprehensible ruleset form. We will summarize the algorithms working in C4.5and conclude with some research issue.

Algorithm

C4.5 is a well-known algorithm used to generate a decision trees. Given a set of cases S, C4.5 first grows an initial tree with the divide-and-conquer algorithm as follows:

If all the cases in S belong to the same class or S is small, the tree is a leaf labeled with the most frequent class in S.

Else, choose a test based on a single attribute with two or more outcomes. Make this test the root of the tree with one branch for each outcome of the test, partition S into corresponding subsets S1, S2, . . . according to the outcome for each case, and apply the same procedure recursively to each subset.

There are generally many tests that could be chosen in this last step. C4.5 uses two heuristic criteria to rank possible tests: information gain, which minimizes the total entropy of the subsets {Si} and the default gain ratio that divides information gain by the information provided by the test outcomes.

Attributes can be either numeric or nominal and this determines the format of the test outcomes. For an attribute B they are {B ≤ t, AB>t} where the threshold t is found by sorting S on the values of B and choosing the split between successive values that maximizes the criterion above.

Disadvantage of C4.5

The main disadvantage of C4.5 is the amount of memory and CPU time. C4.5 require a lot of memory space and CPU time for execution.

The k-means (Introduction)

The k-means algorithm is a simple clusters based data mining algorithms. It is an iterative method to partition a given dataset into a number of clusters. This algorithm has been discovered by several researchers across different disciplines, most notably Lloyd (1957, 1982) [22], Forgey (1965), Friedman and Rubin (1967), and McQueen (1967). A detailed history of k-means along with descriptions of several variations is given in [23]. Gray and Neuhoff [13] provide a nice historical background for k-means placed in the larger context of hill-climbing algorithms.

Algorithm

The algorithm operates on a set of d-dimensional vectors, D = {xi | i = 1, . . . , N}, where xi ∈d denotes the ith data point. The algorithm is initialized by picking k points in d as the initial k cluster representatives or "centroids". Techniques for selecting these initial seeds include sampling at random from the dataset, setting them as the solution of clustering a small subset of the data or perturbing the global mean of the data k times. Then the algorithm iterates between two steps till convergence:

Step 1: Data Assignment. Each data point is assigned to its closest centroid, with ties broken arbitrarily. This results in a partitioning of the data.

Step 2: Relocation of "means". Each cluster representative is relocated to the center (mean) of all data points assigned to it. If the data points come with a probability measure (weights), then the relocation is to the expectations (weighted mean) of the data partitions.

The algorithm converges when the assignments (and hence the cj values) no longer change.

Disadvantage of k-means

k-means is a very simple algorithm and its speed allows it to run on large datasets. Its disadvantage is that it does not give the same result with each run, since the resulting clusters depend on the initial random assignments (the k-means++ algorithm addresses this problem by seeking to choose better starting clusters). It minimizes intra-cluster variance, but does not ensure that the result has a global minimum of variance. Another disadvantage is the requirement for the concept of a mean to be definable which the case is not always. For such datasets the k-medoids variants is appropriate. An alternative, using a different criterion for which points are best assigned to which centre is k-medians clustering.

Support Vector Machines (Introduction)

The support vector machine is a training for learning classification and regression rules from data, for example the SVM can be used to learn polynomial, radial basis function (RBF) and multi-layer perceptron (MLP) classifiers. SVMs were first suggested by Vapnik in the 1960s for classification and have recently become an area of intense research owing to developments in the techniques and theory coupled with extensions to regression and density estimation. SVMs arose from statistical learning theory; the aim being to solve only the problem of interest without solving a more difficult problem as an intermediate step. SVMs are based on the structural risk minimisation principle, closely related to regularisation theory. This principle incorporates capacity control to prevent overfitting and thus is a partial solution to the bias-variance trade-off dilemma 8

In a two-class learning task, the aim of SVM is to find the best classification function to distinguish between members of the two classes in the training data. The metric for the concept of the "best" classification function can be realized geometrically. For a linearly separable dataset, a linear classification function corresponds to a separating hyperplane f (x) that passes through the middle of the two classes, separating the two. Once this function is determined, new data instance xn can be classified by simply testing the sign of the function f (xn); xn belongs to the positive class if f (xn) > 0. Because there are many such linear hyperplanes, what SVM additionally guarantee is thatthe best such function is found by maximizing the margin betwee the two classes.

Disadvantage of SVM

One of the initial drawbacks of SVMis its computational inefficiency. However, this problem is being solved with great success. One approach is to break a large optimization problem into a series of smaller problems, where each problem only involves a couple of carefully chosen variables so that the optimization can be done efficiently. The process iterates until all the decomposed optimization problems are solved successfully. A more recent approach is to consider the problem of learning an SVM as that of finding an approximate minimum enclosing ball of a set of instances. These instances, when mapped to an N-dimensional space, represent a core set that can be used to construct an approximation to the minimum enclosing ball. Solving the SVM learning problem on these core sets can produce a good approximation solution in very fast speed. For example, the core-vector machine [30] thus produced can learn an SVM for millions of data in seconds. Rules with confidence larger than or equal to a user specified minimum confidence.

AdaBoost

Description of the algorithm (Introduction)

AdaBoost, is a machine learning algorithm. formulated by Yoav Freund and Robert Schapire who won the "Gödel Prize" in 2003 for their work. It is a meta learning algorithm. It can be used in combination with many other types of learning algorithms to

improve their performance. AdaBoost is adaptive in the sense that subsequent weak learners are tweaked in favor of those instances misclassified by previous classifiers.

While every learning algorithm will tend to suit some problem types better than others, and will typically have many different parameters and configurations to be adjusted before achieving optimal performance on a dataset, AdaBoost (with decision trees as the weak learners) is often referred to as the best out-of-the-box classifier. When used with decision tree learning, information gathered at each stage of the AdaBoost algorithm about the relative 'hardness' of each training sample is fed into the tree growing algorithm such that later trees tend to focus on harder to classify examples.

In order to deal with multi-class problems, Freund and Schapire presented the Ada-Boost.M1 algorithm [9] which requires that the weak learners are strong enough even on hard distributions generated during the AdaBoost process. Another popular multi-class version of AdaBoost is AdaBoost.MH [26] which works by decomposing multi-class task to a series of binary tasks. AdaBoost algorithms for dealing with regression problems have also been studied. Since many variants of AdaBoost have been developed during the past decade, Boosting has become the most important “family” of ensemble methods.

Conclusion

Data mining is a broad area that integrates techniques from several fields including machine learning, statistics, pattern recognition, artificial intelligence, and database systems, for the analysis of large volumes of data. There have been a large number of data mining algorithms rooted in these fields to perform different data analysis tasks. The algorithms identified on Data Mining (ICDM) and presented in this article are among the most influential algorithms for classification [18, 21], clustering [5,12,14–17], statistical learning [10,28], association analysis [2,6,20,23,27], and link mining. We hope this survey paper can inspire more researchers in data mining to further explore these algorithms, including their impact and new research issues.

References

1. Banerjee A, Merugu S, Dhillon I, Ghosh J (2005) Clustering with Bregman divergences. J Mach LearnRes 6:1705–1749
2. Bonchi F, Lucchese C (2006) On condensed representations of constrained frequent patterns. Knowl Inf Syst 9(2):180–201
3. Breiman L, Friedman JH, Olshen RA, Stone CJ (1984) Classification and regression trees. Wadsworth, Belmont
4. Brin S, Page L (1998) The anatomy of a large-scale hypertextual Web Search Sngine. Comput Networks 30(1–7):107–117
5. Chen JR (2007) Making clustering in delay-vector space meaningful. Knowl Inf Syst 11(3):369–385
6. Chi Y, Wang H, Yu PS, Muntz RR (2006) Catch the moment: maintaining closed frequent it emsets over a data stream sliding window. Knowl Inf Syst 10(3):265–294
7. Dempster AP, Laird NM, Rubin DB (1977) Maximum likelihood from incomplete data via the EM algorithm (with discussion). J Roy Stat Soc B 39:1–38
8. Dietterich TG (1997) Machine learning: Four current directions. AI Mag 18(4):97–136
9. Freund Y, Schapire RE (1997) A decision-theoretic generalization of on-line learning and an application to boosting. J Comput Syst Sci 55(1):119–139
10. Fung G, Stoeckel J (2007) SVM feature selection for classification of SPECT images of Alzheimer’sdisease using spatial information. Knowl INF Syst 11(2):243–258
11. Golub GH, Van Loan CF (1983) Matrix computations. The Johns Hopkins University Press
12. Gondek D, Hofmann T (2007) Non-redundant data clustering. Knowl Inf Syst 12(1):1–24
13. Gray RM, Neuhoff DL (1998) Quantization. IEEE Trans Inform Theory 44(6):2325–2384
14. Hu T, Sung SY (2006) Finding centroid clusterings with entropy-based criteria. Knowl Inf Syst 10(4):505–514
15. Jin R, Goswami A, Agrawal G (2006) Fast and exact out-of-core and distributed *k*-means clustering. Knowl Inf Syst 10(1):17–40
16. Kobayashi M, Aono M (2006) Exploring overlapping clusters using dynamic re-scaling and sampling. Knowl Inf Syst 10(3):295–313
17. Koga H, Ishibashi T, Watanabe T (2007) Fast agglomerative hierarchical clustering algorithm using Locality-Sensitive Hashing. Knowl Inf Syst 12(1):25–53
18. Kukar M (2006) Quality assessment of individual classifications in machine learning and data mining. Knowl INF Syst 9(3):364–384
19. Langville AN, Meyer CD (2006) Google’s PageRank and beyond: the science of search engine rankings. Princeton University Press, Princeton
20. Leung CW-k, Chan SC-f, Chung F-L (2006) A collaborative filtering framework based on fuzzy association rules and multiple-level similarity. Knowl Inf Syst 10(3):357–381
21. Li T, Zhu S, OgiharaM (2006) Using discriminant analysis for multi-class classification: an experimental
22. Lloyd SP (1957) Least squares quantization in PCM. Unpublished Bell Lab. Tech. Note, portions presented at the Institute of Mathematical Statistics Meeting Atlantic City, NJ, September 1957. Also, IEEE Trans Inform Theory (Special Issue on Quantization), vol IT-28, pp 129–137, March 1982
23. Leung CK-S, Khan QI, Li Z, Hoque T (2007) CanTree: a canonical-order tree for incremental frequent pattern mining. Knowl Inf Syst 11(3):287–311
24. Page L, Brin S, Motwami R, Winograd T (1999) The PageRank citation ranking: bringing order to the Web. Technical Report 1999–0120, Computer Science Department, Stanford University

25. 25. Schapire RE, Freund Y, Bartlett P, Lee WS (1998) Boosting the margin: A new explanation for the effectiveness of voting methods. Ann Stat 26(5):1651–1686
26. Schapire RE, Singer Y (1999) Improved boosting algorithms using confidence-rated predictions. Mach Learn 37(3):297–336
27. Steinbach M, Kumar V (2007) Generalizing the notion of confidence. Knowl Inf Syst 12(3):279–299
28. Tao D, Li X, Wu X, Hu W, Maybank SJ (2007) Supervised tensor learning. Knowl Inf Syst 13(1):1–42
29. Toussaint GT (2002) Open problems in geometric methods for instance-based learning JCDCG 273–283
30. Tsang IW, Kwok JT, Cheung P-M (2005) Core vector machines: Fast SVM training on very large data sets. J Mach Learn Res 6:363–392
31. Vapnik V (1995) The nature of statistical learning theory. Springer, New York.