Enhancing K-Means and Naive Bayes for Data Mining

Tarika Verma[1], Dr. Chhavi Rana[2], Monika Boora[3]

[1]Department of Computer Science and Engineering, UIET, M.D. University, Rohtak - 124001, Haryana, India
Email: tarika.verma@yahoo.co.in
[2]Assistant Professor, Department of Computer Science and Engineering, UIET,
M.D. University, Rohtak - 124001, Haryana, India
Email: chhavi1jan@yahoo.com
[3]Department of Computer Science and Engineering, UIET, M.D. University, Rohtak - 124001, Haryana, India
Email: monikaboora93@gmail.com

***Abstract*: Data Mining is used to peruse the recurring trends in the given data sets. This analysis may put on display, new features and insights to the given data set and thus alleviate its usage. This paper includes multiple standard data mining algorithms, K-means clustering models, and naive Baye's classifiers and discusses their overview to cluster the objects belonging to the given data sets. Also, an algorithm is proposed so as to implement the dynamic clustering and to improve the flexibility of these algorithms.**

Keywords: Data Mining Algorithms, Dynamic clustering, Enhancement, k-Means, Naive Bayes

I. INTRODUCTION

The algorithm in data mining (aka machine learning) is used to convert the unorganized data sets into an organized model. For this conversion, the algorithm searches for the patterns which are recognizable or various kind of trends in the given data set and the results are iterated many times to acquire the most favorable parameters required to model them. This model may take any form such as a clusters' set, a decision tree, a mathematically model or simply a set of rules.

These models for the mining can be deployed to various applications, for instance:

1) Forecasting: Sales prediction, prediction of loads or downtime of the server
2) Risk and probability: Choosing the customers for a specific service or product or in the prediction of certain disease by calculating the occurrences of the symptoms or other outcomes and calculating their respective probabilities.
3) Recommendations: Determination of the a-priori products selling, generating recommendations about the products consumed together
4) Finding sequences: The customer interest areas are scrutinized, prognosing its outcomes.
5) Grouping: Separating objects or customers into the cluster of co-related items, analyzing their trends and predicting their relevant interest areas.[7]

The obtained parameters are finally applied on the entire data set so as to excerpt actionable trends and patterns. For further analyzing the pattern, a detailed statistics can be created subsequently.

II. K-MEANS CLUSTERING

This algorithm divides M data points (which are in N dimensions) into K clusters in order to minimize the within-cluster sum of squares. We try to achieve the "local" optima solutions such that no inter-cluster data point manoeuvre reduces the within-cluster sum of squares. [1]

Basically, this algorithm creates k clusters and pairs similar type of objects in a unique cluster. Thus k clusters are formed in such a way that the constituents of a certain cluster are similar as compared to the non-cluster constituents of a certain data set.

A. Proceedings

Initially, k initial cluster centres are selected and then iteratively refined as:

1. Each instance d_i is assigned to its closest cluster centre.
2. Each cluster centre C_j is then updated and this becomes equivalent to the mean of its elemental instances. [2]

These steps are iterated until no further change is there in the apportionment of instances to clusters. Simply we may say that iterations are continued till cluster memberships are stabilized. This is called convergence.

DOI: 10.21817/ijet/2017/v9i3/170903S053

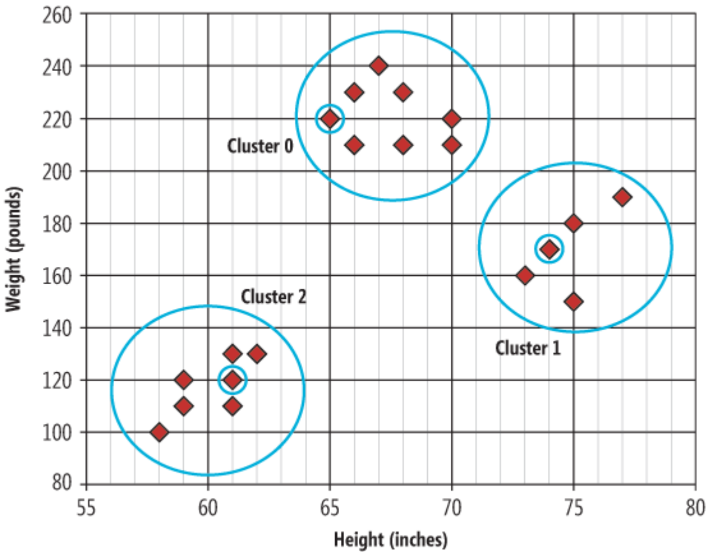

Fig 1. K-means output for sample data set having clustered data and respective centroids [8], k=3

Here, using random instances from the data items, the clusters are initialized and the data items used are either numeric or symbolic, solely. For the numeric type of features, the Euclidean distance metric is used and for the symbolic type of features, the Hamming distance is computed. [3]

B. Advantages

K-means algorithm is simple and easy to implement. It has quirky speed and efficiency as compared to other algorithms.

It could be brought to bear for exploring whether there are overlooked patterns or relationships in the dataset.

C. Limitations

k-Means can only be used to initial clustering of an immense dataset and it is needed to be followed by a big-budget cluster analysis on the parted out sub-clusters.

III. NAIVE BAYES

In this method, the data items are scrutinized statistically for their classification. Cogitating on these inferential statistics the decisions and inferences are puzzled out by applying the Bayesian reasoning. Thus precedent events knowledge guides subsequent similar events.

In this algorithm, it is assumed that every trait of the data being classified is independent of all other traits given the class. This assumption of independent dataset traits grounds it’s naming with the word naive. Bayes’ Theorem is denominated after prominent statistician Thomas Bayes.

It may be represented as:

$$P(h/e) = \frac{P(e/h) . P(h)}{P(e)} = \frac{P(e/h). P(h)}{P(e/h). P(h) + P(e/\sim h). P(\sim h)}$$

Eqn 1: Baye’s Theorem [4]

- Where P(h) is the prior probability of hypothesis h.
- P(h/e) is the posterior probability of hypothesis h (in presence of the evidence e).
- P(e/h) is the likelihood of evidence e on hypothesis h.

A. Proceedings

The elucidated equation for classification may be put in writing as:

$$P(\text{ Class A}| \text{ Trait 1, Trait 2}) = \frac{P(\text{Trait 1}| \text{ Class A}) . P(\text{Trait 2}| \text{ Class A}) . P(\text{Class A})}{P(\text{Trait 1}). P(\text{Trait 2})}$$

Eqn 2:Simplified Baye’s Theorem [4]

This equation evaluates the probability of Class A given that Trait 1 and Trait 2 is prior to the Class A. In other words, if we see Trait 1 and Trait 2, this is the conditional probability of Class A.

Once the frequency tables are calculated, we may classify an unknown data object to a class by calculating the probabilities of its likelihood for all the classes, and then choose the highest probability.

B. Advantage

This model is simple, straightforward, elegant and robust. It is a piece of cake as it just involves calculating the probabilities. It is amidst old classification algorithms, and yet venerable. [6] It is widely used in areas such as text classification and spam filtering.

IV. PROPOSED ALGORITHM

Proposed clustering approach is to classify the data items into a number of groups, which are unknown initially. Presuming the number of groups or clusters is a positive integer, K.

1. Measuring the distance between object and centre of the group (or centroid) does the grouping.
2. Then Naive Bayes may be used to give weights to the classes to which the current object under scrutiny belongs to. Also, a threshold function could be manoeuvred to delay a certain object to wait, if the weight of belongingness to a group or cluster is very less. Thus next object is heeded accordingly. The threshold function allows adding constraints to the belongingness of an item to any cluster
3. The objects are iteratively grouped into the existing clusters using above steps.
4. After each grouping, the clusters-centroids are appraised again as in the K-mean algorithm. This creates the dynamic clustering of the given data items.

It could improve the chances of finding the global optima.

Further, the space requirements may be improved by storing all data in secondary memory and transferring the centroids and the object (under consideration) in the main memory.

V. CONCLUSIONS

The K- means clustering model creates k clusters and pairs similar type of objects in a unique cluster are formed whereas the Naive Bayes Model provides a statistical way to categorize an object. Thus both of these standard algorithms may be implemented together in order to procure a more flexible model for the data mining. Also, further modifications may be done to reduce the complicatedness of the algorithms as stated in the proposed algorithms. This dynamic clustering update may lead to the global optima and forbye local optima problem may be avoided.

REFERENCES

[1] J. A. Hartigan, and M. A. Wong, “A K-Means Clustering Algorithm”, Journal of the Royal Statistical Society. Series C (Applied Statistics), Vol. 28, No. 1 (1979), pp. 100-108

[2] Kiri Wagstaf, and Claire Cardie, “Constrained K-means Clustering with Background Knowledge”, Proceedings of the Eighteenth International Conference on Machine Learning, 2001, p. 577–584.

[3] Raymond Li, “Data Mining Algorithms, Explained”, Tutorials, KDnuggets, 2015

[4] Vapnik, Vladimir Naumovich, and Vlamimir Vapnik. Statistical learning theory. Vol. 1. New York: Wiley, 1998.

[5] Wu, Xindong, et al. "Top 10 algorithms in data mining." Knowledge and information systems 14.1 (2008): 1-37.

[6] Community Library, Microsoft, “Data Mining Concepts”, Microsoft SQL Server, 2016

[7] James McCaffrey, “Data Clustering - Detecting Abnormal Data Using k-Means Clustering”, MSDN Magazine Vol 28 No. 2, February 2013

AUTHOR PROFILE

Tarika Verma is currently pursuing Master in Technology degree at MD University, Rohtak, Haryana, India. She has also got TEQUIP World Bank scholarship for the meritorious students and has been teaching under the same scheme since last two years. She has been interested in the field of Database Management, Data Analytics, and Big Data Analysis. She has attended conferences and also has presented papers related to this field.

Dr. Chhavi Rana Balhara has an experience of over 10 years teaching Data Mining and Web Development subjects at various engineering institutes. She has been interested in the area of Data Mining research from the past 8 years attending around 25 conferences and presenting papers related to this field. Also, she has published 30 papers in reputed journals including Springer and Elsevier. Besides data mining, her research interests also include information management, information retrieval, and ICT. She has supervised 20 M.Tech thesis and currently supervising 4 M.Tech thesis and 4 Ph.D. students. She has also won DST Travel grant twice to present the paper in USA and Spain as well as TEQUIP World Bank travel grant to present a paper in University of Sydney, Australia. She has also been a reviewer on IEEE Transaction’s on Systems, Man and Cybernetics: Systems, Artificial Intelligence Review, Springer as well as Inderscience Publishers. She has also published 4 books on her research work