American Journal of Engineering Research (AJER)
e-ISSN: 2320-0847 p-ISSN : 2320-0936
Volume-5, Issue-11, pp-273-277
www.ajer.org
Research Paper Open Access

Data Mining & It's Algorithms

Ashiqur Rahman[1], Asaduzzaman Noman[2], Hasan Mahmud Mishu[3], Farjana Sumi[4]
[1](M.Sc in Information Technology (IT), Jahangirnagar University, Bangladesh)
[2](CSE, Asian University of Bangladesh, Bangladesh)
[3, 4](MCSE, Royal University of Dhaka, Bangladesh)

ABSTRACT: *Generally, data mining (sometimes called data or knowledge discovery) is the process of analyzing data from different perspectives and summarizing it into useful information - information that can be used to increase revenue, cuts costs, or both. 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 correlations or patterns among dozens of fields in large relational databases. Although data mining is a relatively new term, the technology is not. Companies have used powerful computers to sift through volumes of supermarket scanner data and analyze market research reports for years. However, continuous innovations in computer processing power, disk storage, and statistical software are dramatically increasing the accuracy of analysis while driving down the cost.*
Keywords: *Data Mining, Warehouses, Clusters, Association, Decision Trees.*

I. DATA, INFORMATION & KNOWLEDGE

Data

Data are any facts, numbers, or text that can be processed by a computer. Today, organizations are accumulating vast and growing amounts of data in different formats and different databases. This includes:

- operational or transactional data such as, sales, cost, inventory, payroll, and accounting
- nonoperational data, such as industry sales, forecast data, and macro-economic data
- meta data - data about the data itself, such as logical database design or data dictionary definitions

Information

The patterns, associations, or relationships among all this data can provide information. For example, analysis of retail point of sale transaction data can yield information on which products are selling and when.

Knowledge

Information can be converted into *knowledge* about historical patterns and future trends. For example, summary information on retail supermarket sales can be analyzed in light of promotional efforts to provide knowledge of consumer buying behavior [1]. Thus, a manufacturer or retailer could determine which items are most susceptible to promotional efforts.

Data Warehouses

Dramatic advances in data capture, processing power, data transmission, and storage capabilities are enabling organizations to integrate their various databases into *data warehouses [2]*. Data warehousing is defined as a process of centralized data management and retrieval. Data warehousing, like data mining, is a relatively new term although the concept itself has been around for years. Data warehousing represents an ideal vision of maintaining a central repository of all organizational data. Centralization of data is needed to maximize user access and analysis. Dramatic technological advances are making this vision a reality for many companies. And, equally dramatic advances in data analysis software are allowing users to access this data freely. The data analysis software is what supports data mining.

II. HOW DOES DATA MINING WORK?

While large-scale information technology has been evolving separate transaction and analytical systems, data mining provides the link between the two. Data mining software analyzes relationships and patterns in stored transaction data based on open-ended user queries. Several types of analytical software are available: statistical, machine learning, and neural networks. Generally, any of four types of relationships are sought:

- **Classes**: Stored data is used to locate data in predetermined groups. For example, a restaurant chain could mine customer purchase data to determine when customers visit and what they typically order. This information could be used to increase traffic by having daily specials.
- **Clusters**: Data items are grouped according to logical relationships or consumer preferences. For example, data can be mined to identify market segments or consumer affinities.
- **Associations**: Data can be mined to identify associations. The beer-diaper example is an example of associative mining [3].
- **Sequential patterns**: Data is mined to anticipate behavior patterns and trends. For example, an outdoor equipment retailer could predict the likelihood of a backpack being purchased based on a consumer's purchase of sleeping bags and hiking shoes.

Data mining consists of five major elements:

- Extract, transform, and load transaction data onto the data warehouse system.
- Store and manage the data in a multidimensional database system.
- Provide data access to business analysts and information technology professionals.
- Analyze the data by application software.
- Present the data in a useful format, such as a graph or table.

Different levels of analysis are available:

- **Artificial neural networks**: Non-linear predictive models that learn through training and resemble biological neural networks in structure.
- **Genetic algorithms**: Optimization techniques that use processes such as genetic combination, mutation, and natural selection in a design based on the concepts of natural evolution [4].
- **Decision trees**: Tree-shaped structures that represent sets of decisions. These decisions generate rules for the classification of a dataset. Specific decision tree methods include Classification and Regression Trees (CART) and Chi Square Automatic Interaction Detection (CHAID) [5] . CART and CHAID are decision tree techniques used for classification of a dataset. They provide a set of rules that you can apply to a new (unclassified) dataset to predict which records will have a given outcome. CART segments a dataset by creating 2-way splits while CHAID segments using chi square tests to create multi-way splits. CART typically requires less data preparation than CHAID.
- **Nearest neighbor method**: A technique that classifies each record in a dataset based on a combination of the classes of the *k* record(s) most similar to it in a historical dataset (where *k* 1). Sometimes called the *k*-nearest neighbor technique.
- **Rule induction**: The extraction of useful if-then rules from data based on statistical significance.
- **Data visualization**: The visual interpretation of complex relationships in multidimensional data. Graphics tools are used to illustrate data relationships.

III. ALGORITHM'S

The k-means algorithm is a simple iterative method to partition a given dataset into a userspecified number of clusters, k. This algorithm has been discovered by several researchers across different disciplines, most notably Lloyd (1957, 1982), Forgey (1965), Friedman and Rubin (1967), and McQueen (1967). A detailed history of k-means along with descriptions of several variations are given in. Gray and Neuhoff provide a nice historical background for k-means placed in the larger context of hill-climbing algorithms. The algorithm operates on a set of d-dimensional vectors, D ={x| i = 1,...,N},wherexi∈_didenotes the I th data point. The algorithm is initialized by picking k points in _asthe initial k cluster representatives or "centroids". Techniques for selecting these initial seedsinclude sampling at random from the dataset, setting them as the solution of clustering asmall subset of the data or perturbing the global mean of the data k times. Then the algorithmiterates between two steps till convergence [6].

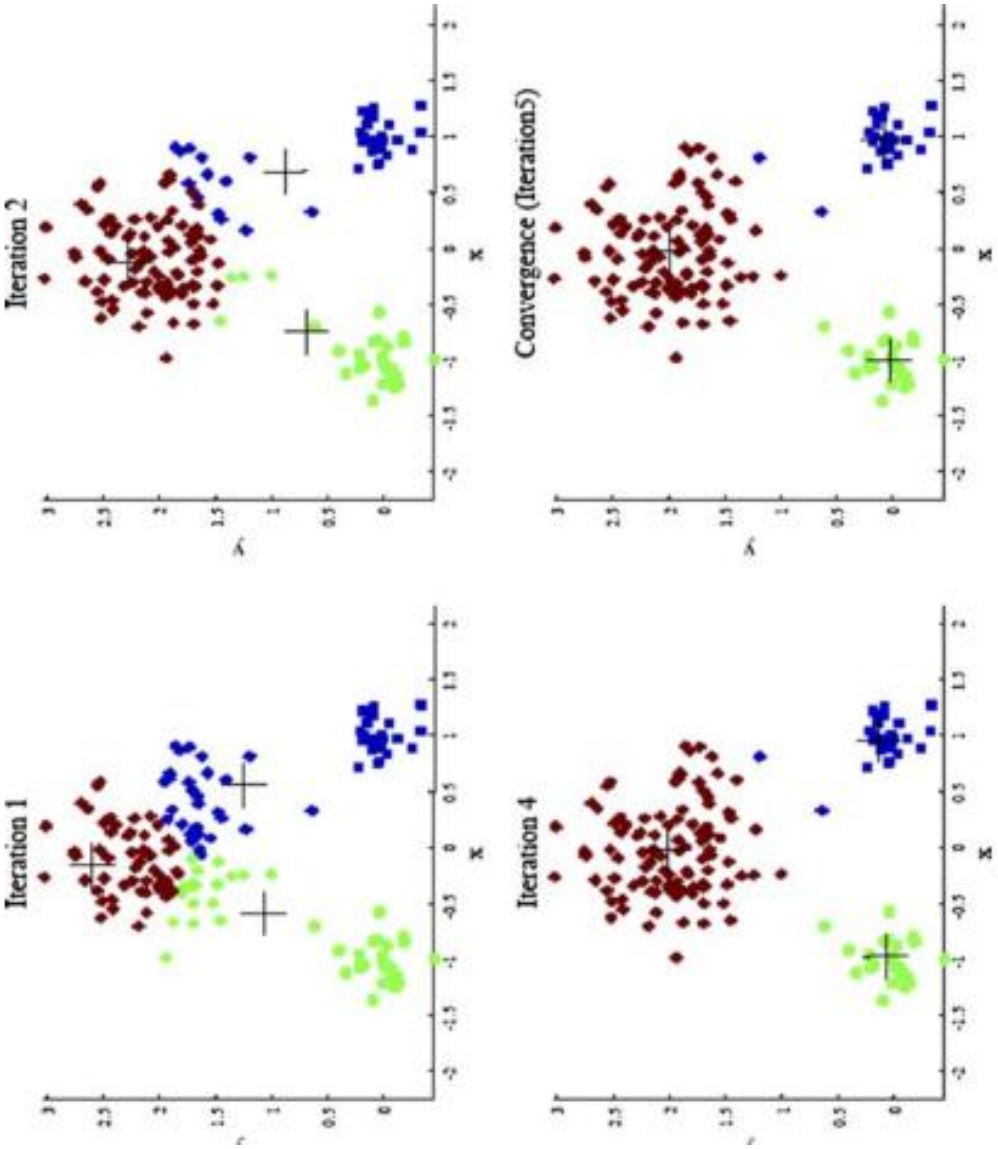
Step 1: Data Assignment. Each data point is assigned to its closest centroid, with ties broken arbitrarily. This results in a partitioning of the data.

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) [7], then the relocation is to the

expectations (weighted mean) of the data partitions. The algorithm converges when the assignments (and hence the values) no longer change [8].The algorithm execution is visually depicted. Note that each iteration needs N × kj comparisons, which determines the time complexity of one iteration. The number of iterations required for convergence varies and may depend on N, but as a first cut, this algorithm can be considered linear in the dataset size. One issue to resolve is how to quantify "closest" in the assignment step. The default measure of closeness is the Euclidean distance, in which case one can readily show that the non-negative cost function,

$$\sum_{i=1}^{N}\left(\underset{j}{\operatorname{argmin}}||\mathbf{x_i}-\mathbf{c_j}||_2^2\right)$$

Will decrease whenever there is a change in the assignment or the relocation steps, and hence convergence is guaranteed in a finite number of iterations. The greedy-descent nature of k-means on a non-convex cost also implies that the convergence is only to a local optimum, and indeed the algorithm is typically quite sensitive to the initial centroid locations. Illustrates how a poorer result is obtained for the same a different choice of the three initial centroids. The local minima problem can be countered to some,



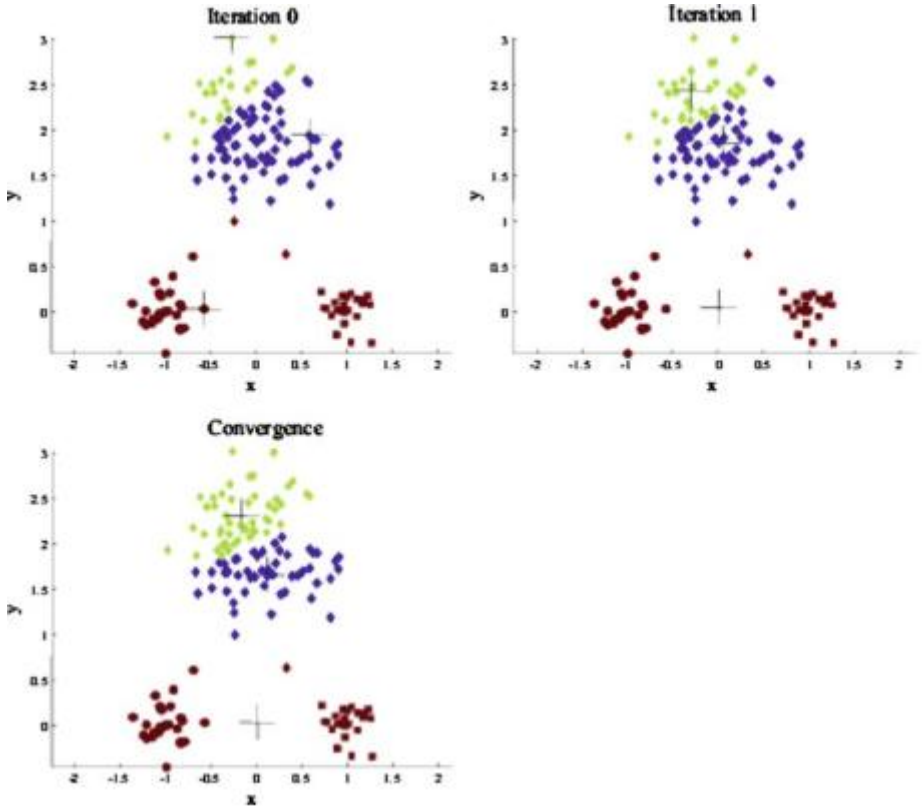


Figure 1& 2: Data Mining Algorithms

IV. LIMITATIONS

In addition to being sensitive to initialization, the k-means algorithm suffers from several other problems. First, observe that k-means is a limiting case of fitting data by a mixture of k Gaussians with identical, isotropic covariance matrices (_ = s2I), when the soft assignments of data points to mixture components are hardened to allocate each data point solely to the most likely component. So, it will falter whenever the data is not well described by reasonably separated spherical balls, for example, if there are non-covex shaped clusters in the data. This problem may be alleviated by rescaling the data to "whiten" it before clustering, or by using a different distance measure that is more appropriate for the dataset. For example, information-theoretic clustering uses the KL-divergence to measure the distance between two data points representing two discrete probability distributions. It has been recently shown that if one measures distance by selecting any member of a very large class of divergences called Bregman divergences during the assignment step and makes no other changes, the essential properties of k-means, including guaranteed convergence, linear separation boundaries and scalability, are retained. This result makes k-means effective for a much larger class of datasets so long as an appropriate divergence is used. K-means can be paired with another algorithm to describe non-convex clusters. One first clusters the data into a large number of groups using k-means. These groups are then agglomerated into larger clusters using single link hierarchical clustering, which can detect complex shapes. This approach also makes the solution less sensitive to initialization, and since the hierarchical method provides results at multiple resolutions, one does not need to pre-specify k either. The cost of the optimal solution decreases with increasing k till it hits zero when the number of clusters equals the number of distinct data-points. This makes it more difficult to (a) directly compare solutions with different numbers of clusters and (b) to find the optimum value of k. If the desired k is not known in advance, one will typically run k-means with different values of k, and then use a suitable criterion to select one of the results. For example, SAS uses the cube-clustering-criterion, while X-means adds a complexity term (which increases with k) to the original cost function and then identifies the k which minimizes this adjusted cost. Alternatively, one can progressively increase the number of clusters, in conjunction with a suitable stopping criterion. Bisecting k-means achieves this by first putting all the data into a single cluster, and then recursively splitting the least compact cluster into two using 2-means. The celebrated LBG algorithm used for vector quantization doubles the number of clusters till a suitable code-book size is obtained. Boththese approaches thus alleviate the need to know k beforehand. The algorithm is also sensitive to the presence of outliers, since "mean" is not a robust statistic. A preprocessing step to remove outliers can be helpful. Post-processing the results, for example to eliminate small clusters, or to merge close clusters into a

large cluster, is also desirable. Ball and Hall's ISODATA algorithm from 1967 effectively used both pre- and post-processing on k-means.

V. 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 10 algorithms identified by the IEEE International Conference on Data Mining (ICDM) and presented inX. Wu et al. this article are among the most influential algorithms for classification, clustering, statistical learning, association analysis and link mining.

REFERENCES

[1]. Agrawal R, Srikant R (1994) Fast algorithms for mining association rules. In: Proceedings of the 20th VLDB conference, pp 487–499
[2]. Ahmed S, Coenen F, Leng PH (2006) Tree-based partitioning of date for association rule mining. Knowl Inf Syst 10(3):315–331
[3]. Banerjee A, Merugu S, Dhillon I, Ghosh J (2005) Clustering with Bregman divergences. J Mach Learn Res 6:1705–1749
[4]. Bezdek JC, Chuah SK, Leep D (1986) Generalized k-nearest neighbor rules. Fuzzy Sets Syst 18(3):237– 256. http://dx.doi.org/10.1016/0165-0114(86)90004-7
[5]. Bloch DA, Olshen RA, Walker MG (2002) Risk estimation for classification trees. J Comput Graph Stat 11:263–288
[6]. Bonchi F, Lucchese C (2006) On condensed representations of constrained frequent patterns. Knowl InfSyst 9(2):180–201
[7]. Breiman L (1968) Probability theory. Addison-Wesley, Reading. Republished (1991) in Classics of mathematics. SIAM, Philadelphia
[8]. Breiman L (1999) Prediction games and arcing classifiers. Neural Comput 11(7):1493–1517