IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661, p- ISSN: 2278-8727Volume 16, Issue 3, Ver. II (May-Jun. 2014), PP 19-24
www.iosrjournals.org

An Insight to Data Stream Mining and other Emerging Learning Algorithms

Shravan Vishwanathan[1], Thirunavukkarasu K [2]
[1]*M.Tech (CSE) Galgotias University, U.P, India*
[2]*Asst Professor, Galgotias University CSE department, U.P India*

Abstract: *Significant advancements have been made in the field of data mining and knowledge discovery. But there is no universal algorithm that can be dynamic enough to extract information from a continuous data stream or from a large stored dataset. A combination of algorithms has to work cohesively in order for the system to yield accurate and timely results according to user specifications. Algorithms are expected to be fast, accurate and visually informative to the user. We have tried to analyse the most effective and recent algorithms and techniques which have been developed to mine information from different data sources. We project our insight on the working of these algorithms and elicit possible loopholes and limitations.*
Keywords : *Data mining, Knowledge discovery, data stream mining, high speed data streams*

I. Introduction

Data mining is the process of discovering useful patterns in large datasets or streams according to user requirements and constraints. It helps us predict future values and trends. Data mining and knowledge discovery are emerging technologies and rigorous research is being done in these sectors. Algorithms are developed for very specific problems. However, even if an algorithm is fast and considerably accurate to produce desired results, it is generally not presented in a way that is very easily understandable to the user. The results although accurate, should be usable to make desired and correct decisions. Decision trees have played a crucial role in artificial intelligence and knowledge discovery process. It is much faster and result oriented and significant modifications and improvements have also been made in the field pertaining to different environments. Neural networks have been well known for their similarity in decision making process as compared to the human brain. It projects nearly accurate results in large, complex and dynamic data sets. The concept of granular neural network for semi supervised leaning [1] has been analysed and discussed in this paper.

The current challenge for researchers today is to extract live and accurate information from online data i.e data streams. There can be various sources which generate massive live data. Data stream mining is also provided as a service by various companies to their customers in order to help them manage their websites, products or services. VFDT [2] (Very Fast Decision Trees) is one of the most fast, effective and accurate algorithm for high rate online data stream mining. It collects extensive amount of examples to build decision trees using the concept of hoeffding bounds. Various modifications and applications have taken place using this algorithm. Semi-supervised learning has also been a prime research area in the field of machine learning. It lies between supervised and unsupervised learning techniques. This technique utilizes a small amount of labelled data from a large pool of unlabelled data in order to learn a problem. This process may require human interference .This leads to considerable improvement in learning accuracy, which may make complete labelled data impracticable.

We also give our insight on the research for Big data which is one of the imperative fields for knowledge discovery today. The extraction of crucial information from an extremely large and dynamic dataset can be complex and time consuming. The data to be mined can range from thousands of terabytes to petabytes and even zettabyte of data. Furthermore, we have discussed the issue of feature extraction [3], which is a subdivision of dimensionality reduction. It helps in reducing the dimensionality of the data in a high dimensional space. It is also required to understand the working of the algorithms when noisy data is fed as input to the system. Does the algorithm perform efficiently and accurately for inconsistent data? We have tried to suggest some techniques and pointed some issues in the algorithms discussed.

II. High Speed Data Streams

A data stream is a continuous inflow of ordered tuples or instances which may refer to some data records. Mining of data streams is frequently desired to be completed in a single pass or in very few or limited number of passes. Generally there is a low probability of the result to be completely accurate. The problem with mining data stream is that a) the stream cannot be stored either in the main memory or in the secondary memory. This is because even though if the rate of the data stream is slow, the main memory or the secondary memory will be flooded eventually. b) there is no exact solution or result for the sample set, so algorithms have to be

designed to calculate nearly accurate results. This is because data stream incurs concept drift which results in statistical changes in the properties of the variable. This may result in lower accuracy in results as time passes.

VFDT(Very Fast Decision Tree learner) system proposed by P domingos *et al* ,(2000) [2] is considered to be one of the most fast and efficient online data learners. Many algorithms have been developed after this algorithm but VFDT and VFDT boot [1] in most cases outperforms other algorithms. The performance and accuracy is considerably better as compared to C4.5 learning system. The algorithm uses the concept of hoeffding bounds which helps the building of decision tree. VFDT is designed as a hybrid decision tree learner which is proficient in utilizing primary and secondary memory according to available system resources. It is adaptable and generates thousands of examples per second and builds decision tree in a constant amount of time and memory space. The system was initially tested on high volume web page requests. Furthermore, D.Panigrahi *et al,* (2010) [4] proposed a novel diversification algorithm to provide a concise, nearly accurate and relevant summary from a large data stream.

Although, it is a theoretical technique but it aims to select apt features from the data stream to overcome the diversification problem. The algorithm is easily scalable and the theoretical analysis results in a 50% approximation ratio. The authors have assumed features of item set as i.i.d. But considering the data stream from a social networking site the data is heavily linked and linked data has an additional attribute of link. On relative analysis, linked data may be suggested to select more accurate features. Diversification algorithm may not be completely suitable for unsupervised learning. Modifying the algorithm to handle linked data can select accurate features, hence producing optimum results.

A system should be capable of controlling the amount processing data from the data stream in order to use recent and apt volume of data to statistically analyse the online system without flooding memory and over utilizing the secondary storage. This lead to the implementation of data windowing models [5] that limit the amount of processed data.

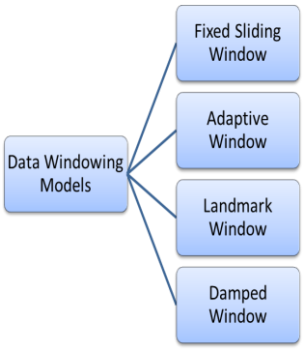

Figure 1: Types of data windowing models

From figure 1 fixed sliding window can be achieved by either two ways, a) include only the recent *n* data points b) project the recent *t* time units of data (*n* and *t* are constants). The size of the window should be accurate because, too narrow window will prove to be accurate but ineffiient in handling noisy data. A wider window will lead to inaccuracy due to the effect of concept drift. Adaptive window is a dynamic model that can adjust the size of the window based on incoming data and the user specified confidence value. The landmark model tracks the course of data point from a a pivot time value. It may not be efficient in data stream analysis but may have limited application with low volume data flow. Damped window model uses the concept of weights which are alloted to data points. The most recent data point is preferred to be allotted with the highest weight to have a greater influence on computation .This eliminates the notion of a boolean decision to include a point or not for analysis.

Kuen Fang Jea *et al,* 2010 [6] proposed a method to extract frequent item sets form a data stream with the help of sliding window protocol using combinatorial approximations. They introduced a novel SWCA (Sliding Window Combinatorial Approximation) and fair-cutter algorithm which helps in dynamic approximation of frequent item sets from a data stream. SWCA conceptually divides the window into segments and handles the sliding window in a segment-based manner. Whereas fair-cutter helps SWCA to dynamically approximate frequent item sets. The algorithm is scalable and the approximation is claimed to achieve high accuracy.

III. Semi-Supervised Learning

Semi-supervised learning is a learning system which has cross characteristics of both supervised (Learning through labelled data) and unsupervised learning (learning through unlabelled data). This learning paradigm uses small amount of labelled data from a vast pool of unlabelled data to learn the system. This

considerably improves the performance of learning accuracy. Guagxia Li *et al*, (2012) [7] quoted the concept of TSVM [8] (Transductive Support Vector Machine) and graph based semi-supervised learning in their article, to improve the quality of a classifier with limited availability of labelled data. Since a two-view TSVM performs 10% better than a single view TSVM, a two-view TSVM is considered in order to improve classification result. A two-view TSVM projects multiple representations of the same data to improve the performance of the classifier. The algorithm performs well when the number of labelled data is small. The article was written for an alternate solution to detect spam product reviews from online forums. However, datasets may not be invariably available in advance. The algorithm discussed may not be suitable for high speed data streams, since the model has to be restarted every time a new data point occurs. Extreme skewness in the data may take more computation time greater computation time.

Interestingly, neural networks have also been one of the most intriguing and complex research areas in data mining. Learning a system through the concept of neurons and treating each node as a function can help in learning complex systems. Nodes are functions in layers that take a variable as input and generate a suitable variable as output that may act as feedback to other layers. Granular neural network [9], [10] is one subsection of neural network which focuses on the granularity of data rather than pure numerical data. It constructs analysable multi- sized local models using fuzzy neurons. The network is adaptable and can handle changes in online environments. Diel Leite *et. al* (2010) [11] introduced the concept of granular neural network for semi-supervised data stream classification. The proposed eGNN (Granular Neural Network) uses the principle of semi-supervised learning and is based on null norm neurons to classify data. eGNN adapts to the learning environment as the system changes. It is capable of classifying data and handling concept drift with considerable efficiency. On performing various computations, neural network does not pick up the relations among the variables in data (example: A data stream from a social networking site). Neural networks tend to be more complex when processing very large data sets. Their accuracy decreases. Example, in order to predict the price of the stock, the network may need numerous inputs including the history of the company prices, current turnover, war conditions, natural calamities and so on. Considering the equation of updating module granularity discussed in the article

$$\rho\ (new) = \rho\left(1 + \frac{\theta}{H_G}\right) \rho(old)$$

ρ is the maximum size that information granules can consume in the feature space. $\{\gamma.....\gamma^{\Theta}\}$ is the set of granules created after a certain number of steps H_G. In case of very large datasets (high speed data streams), H_G may be very high to increase granularity steps. This may resist the system to adapt easily. Since the setup is performed on a small data set, there is little evidence to prove its efficiency in large and complex systems.

IV. Mining Data Sets

Data mining analysis and forecast of variable values on readily available and ordered data is a common practice done in various industries. A dataset is generated or arranged in the form of a table also known as a data matrix. Each column of the data table represents a variable and the each row entry denotes transition values of the variable called as datum. The data set as we know are stored in repositories for reference and statistical analysis. Characteristics of the dataset can be calculated mathematicall0y such as standard deviation and kurtosis. Having discussed the basic definition of a dataset, around 20% web queries are intended for local content. This kind of query retrieval can be satisfied by geo-positional technologies such as GPS and Wi-Fi. Websites like Google, Localpedia, Foursquare, Tripadvisor use spatial keyword query to search local restraunts, hotels, cafe [12],[13],[14].They use this technology in order to return fast and relevant results. Each item is geo tagged with the help of GPS systems. Cellphones, cameras have a dedicated GPS chip which directly tags a photo or a video based on its position in a specific format when desired to be posted online. The content is then segregated and stored in local servers accordingly by the service provider. For e.g, When a common food item like "pizza", "burger" is queried in the search engine the query returns local restraunts nearest to the current location of the user. This example stated is referred as the Top-k spatial query retrieval. Dingming Wu *et. al* (2012) [15] proposed a novel algorithm known as the joint Top-k spatial keyword processing for efficiently processing multiple heavy load queries. It is an index structure for joint processing of Top-k spatial keyword processing. They introduced three algorithms W-IR index structure, W-IBR tree(Inverted Bitmap in Word Partitioning) and GROUP Algorithm. The W-IR algorithm partitions keyword first and then partitions data objects based on spatial locations unlike IR- tree. It can match queries better than IR-tree. Using inverted bitmaps in W-IR tree reduces storage space and saves I/O during query processing. GROUP algorithm on top of W-IBR is the most efficient for processing joint Top-k spatial keyword processing.

Robert J Bayardo *et.al* (2012) [16] introduced two algorithms to find out extermal (maximal and minimal sets) sets from large data sets with the help of minimal storage and processing. The algorithm focuses

on finding extermal sets using cardinality and lexicographic constraints. It is claimed to perform more efficiently to discover frequent item set as compared to GenMax(state of the art frequent item set discoverer)[17]. In this context extermal set is referred with maximal set. Maximal set on the other hand is an independent set that is not a subset of any other complete independent set. Consider it as a set Q such that every edge of a graph has at least one endpoint not in Q and every vertex not in Q has at least one neighbour in Q. On frequently analysing the algorithms are fast and easy to implement. They leverage index structures to reduce algorithm bottlenecks, and item frequency distributions to heuristically minimize the search space and improve locality. Moreover, Single item indexing strategy is specifically used to improve performance. Noticeably, AMS-Lex is significantly faster in finding extermal sets even in graphs. Accordingly the algorithm is tested on multi gigabyte data and not beyond 8GB. What happens for the time and space for extremely large datasets that is petabytes od data? Online data streams may significantly affect the performance of these algorithms, since the cardinality and lexicographic constraints may not always hold true for context drift. Furthermore, AMS-card and AMS-Lex need to scan the input twice to check if the data is memory resident. The algorithm may be improvised to process datasets in one pass. Memory management for extremely large datasets may be considered.

In the year 2010 Toon Calders *et. al* [18] introduced an innovative technique to approximate frequentness probability (of itemsets in large datasets) using central limit theorem [19]. Central limit theorem basically states that the defined mean and variance of large iterations of randomly distributed variables would be normally distributed. The technique is claimed to have much lower computation and overhead time than existing algorithms. Beyond that, an apriori framework has been used to approximate the frequentness probability. The project has been designed in GNU/Linux. There are some queries that have led to ambiguity in the framework such as criteria for sampling data from datasets could not be understood clearly (random samples taken). Wouldn't the computational time increase if the no. of samples exceed? (Taking random sample case) even if it leads to a better normal distribution compromising the processing time. The adjustment of minsup (minsupport) is not clearly specified in the framework.

V. Big Data

We are living in an age where we are bombarded with data from numerous sources. The term Big data [20] has been coined to exhibit the size of data, industries are dealing presently. The data sets generated are so massive and complex that existing database management tools find it difficult to process this amount. The size of the data sets have crossed the range of petabytes and as of from 2012 analysis suggest that industries are dealing with data sets in the order of exabytes. Moreover, it is still growing with real time data generated by wireless sensor networks, mobile devices, cameras, social networking sites, search engines. Imagine, the kind of data generated in real time. As of 2012, about $2.6x10^{18}$bytes of data were created every day. In order to cope up to manage and analyse such quantity, researchers are continuously discovering new methods for an appropriate solution for specific requirements. Interesting open source projects like Apache Hadoop, Spark, MapReduce [21]-[25] are available for Big data processing and distributed file handling. Other than that, contributions are being made in the field by some researchers.

U Kang *et.al* (2012) [26] proposed a Big graph mining system Pegasus. The framework is built on top of MapReduce(Hadoop) to provide the edge of distributed processing. Pegasus is claimed to perform 9.2X – 76X faster than its counterparts. Petabytes of data can be processed efficiently by this system. Basically 5 algorithms have been selected and designed to work in coalescence, Page Rank, Random Walk with Restart(RWR),diameter/radius estimation ,connected components, eigen solver. The system is developed to find patterns and anomalies in real world graphs. Testing the system with Yahoo web(2002 WWW crawler) and Twitter Who- follows-who graph(November 2009) gave a projection of anomalies in connected components. However, having used Hadoop as a base to develop the framework there may be some issues that may raise some concerns. Hadoop breaks data into pieces for store and query. Organizations may not detect patterns and learn insights from data due to undetected links within the data. Also, Hadoop is too slow for companies requiring query response in split second and costly to have a distributed set up.

VI. Feature Extraction

An Abundant quantity of digital information is generated every day in real time with the help of mobility devices and high speed data processing systems. Having it discussed earlier, searching algorithms developed earlier are incompetent to return useful information to the user. One solution is to project or map patterns from an intricate information space into a manageable feature space. Raw data is split into data items that have a frequent pattern. The data items are then mapped to the feature space with the help of a specially designed function. For this reason the mapping of data items to the feature space is termed as feature extraction [27] e.g automatic indexing of textual data and image recognition. Feature extraction in case of low quality data

can prove to be more complicated because of missing and erroneous values, the features extracted may not be accurate and may lead to undesired statistics.

Brian Quanz et.al (2012) [28] presented a technique for knowledge extraction on low quality data. The authors have proposed a technique based on sparse coding, which essentially attempts to find an embedding for the data by assigning feature values based on subspace cluster membership [29]. The current sparse coding is modified for different source and data distributions, to improve its shortcoming for better feature extraction. The system combines two algorithms which are sparse coding with regularization and weighted loss sparse coding. Data with no or little truth information can be utilized to extract information with the help of given auxiliary databases. Extraction of higher features can be discovered with the help of modified sparse code. The article also addresses the issue of knowledge transfer in case of heterogeneous data sources.

VII. Conclusion

This paper gives an insight of the trending and emerging technologies that are critical to manage the ever exploding digital content around the globe. Frameworks and algorithms are being researched rigorously, for revealing patterns and information statistically from petabytes and even exabytes of data. We provide an overview of the most favourable issues that are being handled in the field of knowledge discovery whether be the subdivision of computer science, bio-chemistry or genomes. Feature extraction has become an important area for research and is being applied in image recognition, video frame matching and other areas. Semi-supervised learning environment is considered to be more efficient to learn from very few labelled data, this can be prove to perform well in case of certain amount of noise in the data. Big data is causing a concern in matters of physical storage, analysis and security and solutions are discovered for the issue. Data streams are ever flowing digitally from numerous sources and dynamic flexibility and efficiency for statistical analysis for current and future trends is expected from a framework. However, there are very few guarantees for 100% accuracy and efficiency from any framework or algorithm.

References

[1]. *Semi-supervised learning*. Vol. 2. Cambridge: MIT press, 2006.
[2]. Domingos, Pedro, and Geoff Hulten. "Mining high-speed data streams." *Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2000.
[3]. Guyon, Isabelle, ed. *Feature extraction: foundations and applications*. Vol. 207. Springer, 2006.
[4]. Panigrahi, Debmalya, et al. "Online selection of diverse results." *Proceedings of the fifth ACM international conference on Web search and data mining*. ACM, 2012.
[5]. Matysiak, Martin. "Data Stream Mining." (2012).
[6]. Jea, Kuen-Fang, and Chao-Wei Li. "A sliding-window based adaptive approximating method to discover recent frequent itemsets from data streams." *Proceedings of the International MultiConference of Engineers and Computer Scientists*. Vol. 1. 2010.
[7]. Li, Guangxia, Kuiyu Chang, and S. Hoi. "Multi-View Semi-Supervised Learning with Consensus." (2012): 1-1.
[8]. Joachims, Thorsten. "Transductive support vector machines." *Chapelle et al.(2006)* (2006): 105-118.
[9]. Zhang, Yan-Qing, et al. "Granular neural networks for numerical-linguistic data fusion and knowledge discovery." *Neural Networks, IEEE Transactions on* 11.3 (2000): 658-667.
[10]. Pedrycz, Witold, and George Vukovich. "Granular neural networks." *Neurocomputing* 36.1 (2001): 205-224.
[11]. Leite, Daniel, P. Costa, and Fernando Gomide. "Evolving granular neural network for semi-supervised data stream classification." *Neural Networks (IJCNN), The 2010 International Joint Conference on*. IEEE, 2010.
[12]. De Felipe, Ian, Vagelis Hristidis, and Naphtali Rishe. "Keyword search on spatial databases." *Data Engineering, 2008. ICDE 2008. IEEE 24th International Conference on*. IEEE, 2008.
[13]. Cao, Xin, et al. "Collective spatial keyword querying." *Proceedings of the 2011 ACM SIGMOD International Conference on Management of data*. ACM, 2011.
[14]. Rocha-Junior, Joao B., et al. "Efficient processing of top-k spatial keyword queries." *Advances in Spatial and Temporal Databases*. Springer Berlin Heidelberg, 2011. 205-222.
[15]. Wu, Dingming, et al. "Joint top-k spatial keyword query processing." *Knowledge and Data Engineering, IEEE Transactions on* 24.10 (2012): 1889-1903.
[16]. Bayardo, Roberto J., and Biswanath Panda. "Fast Algorithms for Finding Extremal Sets." *SDM*. 2011.
[17]. Gouda, Karam, and Mohammed J. Zaki. "Genmax: An efficient algorithm for mining maximal frequent itemsets." *Data Mining and Knowledge Discovery* 11.3 (2005): 223-242.
[18]. Calders, Toon, Calin Garboni, and Bart Goethals. "Approximation of frequentness probability of itemsets in uncertain data." *Data Mining (ICDM), 2010 IEEE 10th International Conference on*. IEEE, 2010.
[19]. Hoffmann-Jørgensen, Jørgen, and G. Pisier. "The law of large numbers and the central limit theorem in Banach spaces." *The Annals of Probability* (1976): 587-599.
[20]. Zikopoulos, Paul, and Chris Eaton. *Understanding big data: Analytics for enterprise class hadoop and streaming data*. McGraw-Hill Osborne Media, 2011.
[21]. Borthakur, Dhruba. "The hadoop distributed file system: Architecture and design." (2007).
[22]. Zaharia, Matei, et al. "Spark: cluster computing with working sets." *Proceedings of the 2nd USENIX conference on Hot topics in cloud computing*. 2010.
[23]. Dean, Jeffrey, and Sanjay Ghemawat. "MapReduce: simplified data processing on large clusters." *Communications of the ACM* 51.1 (2008): 107-113.
[24]. Chu, Cheng, et al. "Map-reduce for machine learning on multicore." *Advances in neural information processing systems* 19 (2007): 281.

[25]. Shvachko, Konstantin, et al. "The hadoop distributed file system." *Mass Storage Systems and Technologies (MSST), 2010 IEEE 26th Symposium on*. IEEE, 2010.
[26]. Kang, U., and Christos Faloutsos. "Big graph mining: algorithms and discoveries." *ACM SIGKDD Explorations Newsletter* 14.2 (2013): 29-36.
[27]. Liu, Liu Huan, and Hiroshi Motoda, eds. *Feature extraction, construction and selection: A data mining perspective*. Springer, 1998.
[28]. Quanz, Brian, Jun Luke Huan, and Meenakshi Mishra. "Knowledge transfer with low-quality data: a feature extraction issue." *Knowledge and Data Engineering, IEEE Transactions on* 24.10 (2012): 1789-1802.
[29]. Parsons, Lance, Ehtesham Haque, and Huan Liu. "Subspace clustering for high dimensional data: a review." *ACM SIGKDD Explorations Newsletter* 6.1 (2004): 90-105.