Volume 6, Issue 12, December 2016 ISSN: 2277 128X

International Journal of Advanced Research in Computer Science and Software Engineering

Research Paper

Available online at: www.ijarcsse.com

An Efficient Clustering Algorithm for Outlier Detection in Data Streams

G. R. Thippeswamy
Professor, Department of Computer Science,
Sri Revena Siddeswara Institute of Technology,
Bangalore, Karnataka, India

Dr. K. Krishnamoorthy
Professor, Department of Computer Science and Engg.,
Shanmuganathan Engineering College,
Tamilnadu, India

Abstract: Information Stream mining has picked up fascination from numerous specialists as there is have to mine substantial dataset which posture diverse difficulties for analysts. Stream information is distinctive contrasted with ordinary information as they are constantly created from various applications which force diverse difficulties like huge, unbounded, idea float for preparing. An item that does not comply with the conduct of typical information article is called exceptions. Anomaly location is utilized as a part of various applications such as extortion location, interruption identification, track natural changes, medicinal determination so there is have to recognize exceptions from information streams. Different methodologies are utilized for anomaly location. Some of them use K-Means calculation for anomaly location in information streams which make a comparative gathering or bunch of information focuses. Information stream bunching systems are profoundly useful to bunch comparative information things in information streams furthermore to distinguish the anomalies from them, so they are called bunch based exception discovery. K-implies calculation is allotment based calculation which is utilized for bunching datasets into number of groups. It is most regular and well known calculation for bunching because of its effortlessness and effectiveness.

Keywords: K-Means, dataset, mining, K-implies.

I. INTRODUCTION

Data is generated from the different applications as number of user is increasing day by day. Data is generated and stored in database which is increasing at fast rate due to technology and hardware improvements. For example humans store different types of data like documents, images, songs, movies, scientific data and many other data into database. There is need to find meaningful information in the form of useful patterns, association, relationships among these data because these large database may contain both useful data and non-useful data. Data mining is the process of mining meaningful information, discovering new patterns, identifying relationship among data etc. from databases, flat files, spatial database, data repositories, temporal database, data stream, transactional database and World Wide Web. Data mining is the process of analyzing the data from different perspective and summarizing it into useful information. Different applications like internet traffic, communication network data, sensor network data, online banking transaction, scientific data social data like emails, web click streams generates infinite volume of data in continuous and incremental manner which is called as data stream. Data stream is different than traditional data as data stream is having characteristics of being evolutionary in nature, massive, fast changing and potentially infinite. Traditional data processing methods don't work well with data streams so there is need to use different techniques for processing stream data from databases, flat files, spatial database, data repositories and temporal Database.

II. SYSTEM ANALYSIS

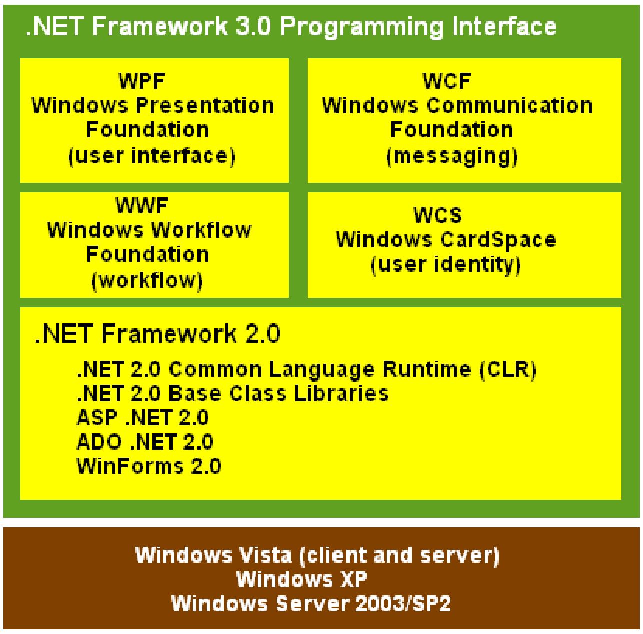
- Outliers are the data points which shows significant diversion from other data points or which is different from the regular or normal data points.Outlierdata points shows different behavior than expected behavioror samebehaviour as other data points. They are generated due to different reasons like malicious activity in network, instrumental error, environmental changes, and errors by human.
- Outliers are the data point which is different and isolated individual data points with respect to all other data points in data set.
- Outliers are the point which is isolated from other data points in the same context. Context of data point refers to semantic relationship among data points.
- Outliers are a subset or a group of data points which appears as outliers with respect to entire dataset.

III. PROPOSED SYSTEM

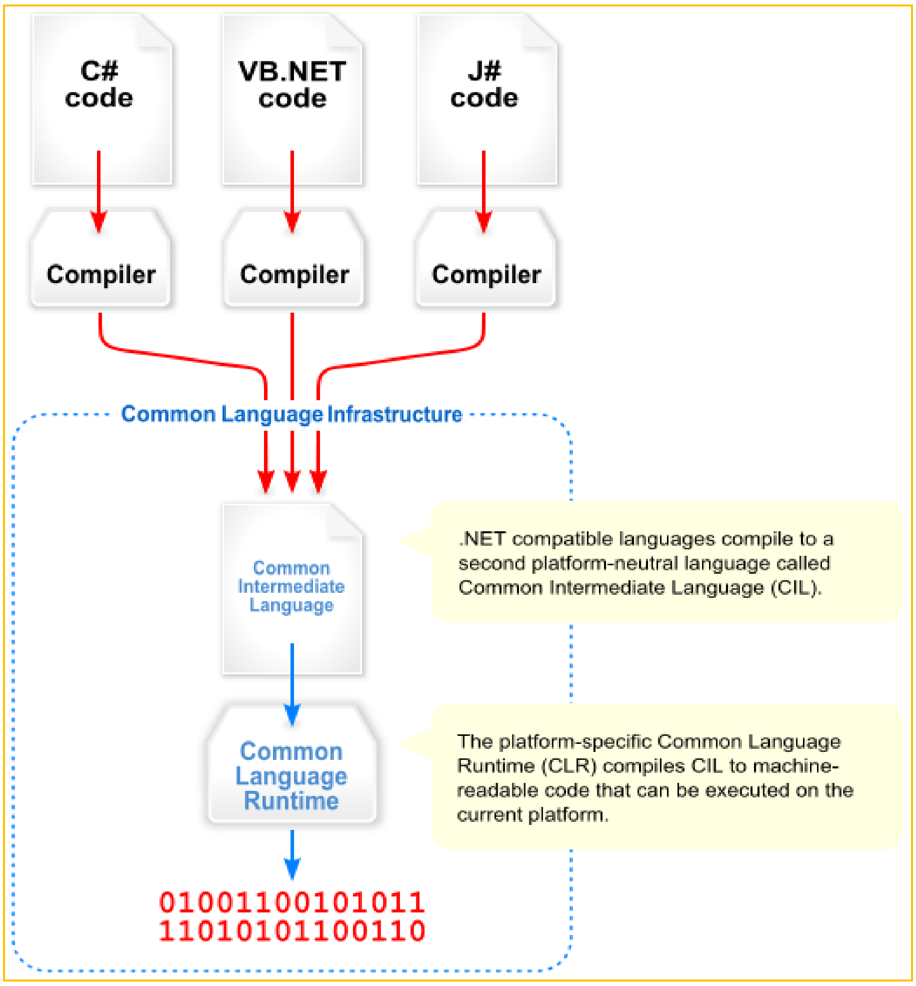
- It is simple and efficient algorithm for clustering dataset. It takes number of cluster k as input parameter and partition a dataset which contains n objects into k clusters.
- An object o in one cluster is similar to objects belong in the same cluster and is called as intracluster similarity. An object o of one cluster is dissimilar with the objects of other cluster called as intercluster similarity.

- First step is to randomly select k numbers of object from the dataset which is used for initially representing centers of k clusters.
- After that for each objects that are not assigned to cluster, an object is assigned to any one cluster based on the similarity with that cluster, based on the distance between cluster mean and object.
- After assigning objects to clusters new mean is computed for each cluster.
- This process is repeats till the criteria function converge.

IV. BLOCK DIAGRAM


.NET Framework 3.0 Programming Interface
WPF
Windows Presentation Foundation
(user interface)
WCF
Windows Communication Foundation
(messaging)
WWF
Windows Workflow Foundation
(workflow)
WCS
Windows CardSpace
(user identity)
.NET Framework 2.0
.NET 2.0 Common Language Runtime (CLR)
.NET 2.0 Base Class Libraries
ASP .NET 2.0
ADO .NET 2.0
WinForms 2.0
Windows Vista (client and server)
Windows XP
Windows Server 2003/SP2

Architecture


C#
code
VB.NET
code
J#
code
Compiler
Compiler
Compiler
Common Language Infrastructure
Common
Intermediate
Language
.NET compatible languages compile to a second platform-neutral language called Common Intermediate Language (CIL).
Common
Language
Runtime
The platform-specific Common Language Runtime (CLR) compiles CIL to machine-readable code that can be executed on the current platform.
01001100101011
11010101100110

Overview of the Common Language Infrastructure

V. COMMON LANGUAGE INFRASTRUCTURE (CLI)

The purpose of the Common Language Infrastructure (CLI) is to provide a language-neutral platform for application development and execution, including functions for exception handling, garbage collection, security, and interoperability. By implementing the core aspects of the .NET Framework within the scope of the CLI, this functionality will not be tied to a single language but will be available across the many languages supported by the framework. Microsoft's implementation of the CLI is called the Common Language Runtime, or CLR.

Memory Management

The .NET Framework CLR frees the developer from the burden of managing memory (allocating and freeing up when done); it handles memory management itself by detecting when memory can be safely freed. Instantiations of .NET types (objects) are allocated from the managed heap; a pool of memory managed by the CLR.. When there is no reference to an object, and it cannot be reached or used, it becomes garbage, eligible for collection. NET Framework includes a garbage collector which runs periodically, on a separate thread from the application's thread, that enumerates all the unusable objects and reclaims the memory allocated to them.

System Design

Systems design is the process of defining the architecture, components, modules, interfaces, and data for a system to satisfy specified requirements. Systems design could be seen as the application of systems theory to product development. There is some overlap with the disciplines of systems analysis, systems architecture and systems engineering.

Input Design is the process converting a user oriented description of the inputs to a computer-based business system into a programmer-oriented specification.

- Input data were found to be available for establishing and maintaining master and transaction files and for creating output records
- The most suitable types of input media for either off-line or on-line devices, where selected after a study of alternative data capture techniques.

VI. CONCLUSION

Outlier detection is the process of finding data objects with behaviors that are very different from expectation. Such objects are called outliers or anomalies. Outlier detection is important in many applications in addition to fraud detection such as medical care, public safety and security, industry damage detection, image processing, sensor/video network surveillance and intrusion detection. Intrusion detection systems (IDS) have become important security tools applied in many contemporary network environments. They gather and analyze information from various sources on hosts and networks in order to identify suspicious activities and generate alerts for an operator.

REFERENCES

[1] Jin, W., Tung, K.H. and Han (2001). Mining top-n local outliers in large DatabasesIn Proc.2001 ACMSIGKDD Int. Conf. Knowledge Discovery in Databases (KDD'01), pp. 293–298,San CA, Aug. 2001.

[2] Babuand widomJ., (2001). Continuous queries over data streams.SIGMOD Record, 30:109–120.

[3] Charu C. Aggarwal, Philip S. Yu, (2001). Outlier detection for high dimensional data, Proc. of the 2001 ACM SIGMOD int. conf. on Management of data, p.37-46, May 21-24, 2001, Santa Barbara, California, United States

[4] Babcock, B., Babu, S. ,Datar, M. , Motwani, R., &Widom,J. (2002).. Principles of Database Systems (PODS'02), pages 1–16, Madison, WI, June 2002.

[5] Gibbons, P.B., &Matias (1998). New sampling-based summary statistics for improving approximate query answers. In Proc. 1998 ACM-SIGMOD ConfManagementofData (SIGMOD'98), pages 331–342, Seattle,WA, June 1998.

[6] Knorr, E., & Ng, R, (1997). A unified notion of outliers: Properties and computation. In Proc. 1997 Int. Conf. Knowledge Discovery and Data Mining (KDD'97), pp. 219–222, Newport Beach, CA, Aug. 1997.

[7] Chandola, V., Banerjee, A., & Kumar, V., (2009). Anomaly detection: A survey. ACM Computing Surveys, 41:1–58.

[8] ChandraSekaran, S., & Franklin, M., (2002). Streaming queries over streamingdata.In Proc.2002 Int. Conf. Very Large Data Bases (VLDB'08)

[9] Babcock, B., Babu, S., Datar, M., Motwani, R. &Widom, J. (2002). Models and issues in data stream systems. In Proc. 2002 ACM Symp. Principles of Database Systems (PODS'02), pages 1–16, Madison, WI, June 2002.

[10] Muthukrishnan, S. (2003). Data streams: algorithms and applications. In Proc. 2003 Annual ACMSIAM Symp. Discrete Algorithms (SODA'03), pages 413–413, Baltimore, MD,Jan. 2003.