Title: Concept drift

URL: https://en.wikipedia.org/wiki/Concept_drift

PageID: 3118600

Categories: Category:Data analysis, Category:Data mining, Category:Machine learning

-----

In predictive analytics , data science , machine learning and related fields, concept drift or drift is an evolution of data that invalidates the data model . It happens when the statistical properties of the target variable, which the model is trying to predict, change over time in unforeseen ways. This causes problems because the predictions become less accurate as time passes. Drift detection and drift adaptation are of paramount importance in the fields that involve dynamically changing data and data models.

Predictive model decay

In machine learning and predictive analytics this drift phenomenon is called concept drift. In machine learning, a common element of a data model are the statistical properties, such as probability distribution of the actual data. If they deviate from the statistical properties of the training data set , then the learned predictions may become invalid, if the drift is not addressed.

Data configuration decay

Another important area is software engineering , where three types of data drift affecting data fidelity may be recognized. Changes in the software environment ("infrastructure drift") may invalidate software infrastructure configuration. "Structural drift" happens when the data schema changes, which may invalidate databases. "Semantic drift" is changes in the meaning of data while the structure does not change. In many cases this may happen in complicated applications when many independent developers introduce changes without proper awareness of the effects of their changes in other areas of the software system.

For many application systems, the nature of data on which they operate are subject to changes for various reasons, e.g., due to changes in business model, system updates, or switching the platform on which the system operates.

In the case of cloud computing , infrastructure drift that may affect the applications running on cloud may be caused by the updates of cloud software.

There are several types of detrimental effects of data drift on data fidelity. Data corrosion is passing the drifted data into the system undetected. Data loss happens when valid data are ignored due to non-conformance with the applied schema. Squandering is the phenomenon when new data fields are introduced upstream the data processing pipeline, but somewhere downstream there data fields are absent.

Inconsistent data

"Data drift" may refer to the phenomenon when database records fail to match the real-world data due to the changes in the latter over time. This is a common problem with databases involving people, such as customers, employees, citizens, residents, etc. Human data drift may be caused by unrecorded changes in personal data, such as place of residence or name, as well as due to errors during data input.

"Data drift" may also refer to inconsistency of data elements between several replicas of a database. The reasons can be difficult to identify. A simple drift detection is to run checksum regularly. However the remedy may be not so easy.

Examples

The behavior of the customers in an online shop may change over time. For example, if weekly merchandise sales are to be predicted, and a predictive model has been developed that works satisfactorily. The model may use inputs such as the amount of money spent on advertising , promotions being run, and other metrics that may affect sales. The model is likely to become less and less accurate over time – this is concept drift. In the merchandise sales application, one reason for concept drift may be seasonality, which means that shopping behavior changes seasonally. Perhaps there will be higher sales in the winter holiday season than during the summer, for example. Concept drift generally occurs when the covariates that comprise the data set begin to explain the variation of your target set less accurately — there may be some confounding variables that have emerged, and that one simply cannot account for, which renders the model accuracy to progressively decrease with time. Generally, it is advised to perform health checks as part of the post-production analysis and to re-train the model with new assumptions upon signs of concept drift.

Possible remedies

To prevent deterioration in prediction accuracy because of concept drift, reactive and tracking solutions can be adopted. Reactive solutions retrain the model in reaction to a triggering mechanism, such as a change-detection test or control charts from statistical process control , to explicitly detect concept drift as a change in the statistics of the data-generating process. When concept drift is detected, the current model is no longer up-to-date and must be replaced by a new one to restore prediction accuracy. A shortcoming of reactive approaches is that performance may decay until the change is detected. Tracking solutions seek to track the changes in the concept by continually updating the model. Methods for achieving this include online machine learning , frequent retraining on the most recently observed samples, and maintaining an ensemble of classifiers where one new classifier is trained on the most recent batch of examples and replaces the oldest classifier in the ensemble.

Contextual information, when available, can be used to better explain the causes of the concept drift: for instance, in the sales prediction application, concept drift might be compensated by adding information about the season to the model. By providing information about the time of the year, the rate of deterioration of your model is likely to decrease, but concept drift is unlikely to be eliminated altogether. This is because actual shopping behavior does not follow any static, finite model . New factors may arise at any time that influence shopping behavior, the influence of the known factors or their interactions may change.

Concept drift cannot be avoided for complex phenomena that are not governed by fixed laws of nature . All processes that arise from human activity, such as socioeconomic processes, and biological processes are likely to experience concept drift. Therefore, periodic retraining, also known as refreshing, of any model is necessary.

See also

Data stream mining

Data mining

Snyk , a company whose portfolio includes drift detection in software applications

Further reading

Many papers have been published describing algorithms for concept drift detection. Only reviews, surveys and overviews are here:

Reviews

Souza, V.M.A.; Reis, D.M.; Maletzke, A.G.; Batista, G.E.A.P.A. (2020). "Challenges in Benchmarking Stream Learning Algorithms with Real-world Data" . Data Mining and Knowledge Discovery . 34 (6): 1805– 58. arXiv : 2005.00113 . doi : 10.1007/s10618-020-00698-5 . S2CID 218470010 .

Krawczyk, B.; Minku, L.L.; Gama, J.; Stefanowski, J.; Wozniak, M. (2017). "Ensemble Learning for Data Stream Analysis: a survey" . Information Fusion . 37 : 132– 156. doi : 10.1016/j.inffus.2017.02.004 . hdl : 2381/39321 . S2CID 1372281 .

Dal Pozzolo, A.; Boracchi, G.; Caelen, O.; Alippi, C.; Bontempi, G. (2015). "Credit card fraud detection and concept-drift adaptation with delayed supervised information" (PDF) . 2015 International Joint Conference on Neural Networks (IJCNN) . IEEE. pp. 1– 8. doi : 10.1109/IJCNN.2015.7280527 . ISBN 978-1-4799-1960-4 . S2CID 3947699 .

Alippi, C. (2014). "Learning in Nonstationary and Evolving Environments" . Intelligence for Embedded Systems . Springer. pp. 211– 247. doi : 10.1007/978-3-319-05278-6_9 . ISBN 978-3-319-05278-6 .

Gama, J.; Žliobait■, I.; Bifet, A.; Pechenizkiy, M.; Bouchachia, A. (1 March 2014), "A survey on concept drift adaptation" (PDF) , ACM Computing Surveys , 46 (4): 1– 37, doi : 10.1145/2523813 , ISSN 0360-0300 , Zbl 1305.68141 , Wikidata Q58204632

Alippi, C.; Polikar, R. (January 2014). "Guest Editorial Learning in Nonstationary and Evolving Environments" . IEEE Transactions on Neural Networks and Learning Systems . 25 (1): 9– 11. doi : 10.1109/TNNLS.2013.2283547 . PMID 24806640 . S2CID 16547472 .

Dal Pozzolo, A.; Caelen, O.; Le Borgne, Y.A.; Waterschoot, S.; Bontempi, G. (2014). "Learned lessons in credit card fraud detection from a practitioner perspective" (PDF) . Expert Systems with Applications . 41 (10): 4915– 28. doi : 10.1016/j.eswa.2014.02.026 . S2CID 12656644 .

Jiang, J. (2008). "A Literature Survey on Domain Adaptation of Statistical Classifiers" (PDF) . School of Computing and Information Systems, Singapore Management University.

Kuncheva, L.I. (2008). "Classifier ensembles for detecting concept change in streaming data: Overview and perspectives" (PDF) . Proceedings of the 2nd Workshop SUEMA 2008 (ECAI 2008) .

Gaber, M.M.; Zaslavsky, A.; Krishnaswamy, S. (June 2005). "Mining Data Streams: A Review" (PDF) . ACM SIGMOD Record . 34 (2): 18– 26. doi : 10.1145/1083784.1083789 . S2CID 705946 .

Kuncheva, L.I. (2004). "Classifier ensembles for changing environments" (PDF) . Multiple Classifier Systems. MCS 2004 . Lecture Notes in Computer Science. Vol. 3077. Springer. pp. 1– 15. doi : 10.1007/978-3-540-25966-4_1 . ISBN 978-3-540-25966-4 .

Tsymbal, A. (2004). The problem of concept drift: Definitions and related work (PDF) (Technical report). Dublin, Ireland: Department of Computer Science, Trinity College. TCD-CS-2004-15.

External links

Software

Frouros : An open-source Python library for drift detection in machine learning systems.

NannyML : An open-source Python library for detecting univariate and multivariate distribution drift and estimating machine learning model performance without ground truth labels.

RapidMiner : Formerly Yet Another Learning Environment (YALE): free open-source software for knowledge discovery, data mining, and machine learning also featuring data stream mining, learning time-varying concepts, and tracking drifting concept. It is used in combination with its data stream mining plugin (formerly concept drift plugin).

EDDM (Early Drift Detection Method): free open-source implementation of drift detection methods in Weka .

MOA (Massive Online Analysis) : free open-source software specific for mining data streams with concept drift. It contains a prequential evaluation method, the EDDM concept drift methods, a reader of ARFF real datasets, and artificial stream generators as SEA concepts, STAGGER, rotating hyperplane, random tree, and random radius based functions. MOA supports bi-directional interaction with Weka .

Datasets

Real

USP Data Stream Repository , 27 real-world stream datasets with concept drift compiled by Souza et al. (2020). Access

Airline , approximately 116 million flight arrival and departure records (cleaned and sorted) compiled by E. Ikonomovska. Reference: Data Expo 2009 Competition [1] . Access

Chess.com (online games) and Luxembourg (social survey) datasets compiled by I. Zliobaite. Access

ECUE spam 2 datasets each consisting of more than 10,000 emails collected over a period of approximately 2 years by an individual. Access from S.J.Delany webpage

Elec2 , electricity demand, 2 classes, 45,312 instances. Reference: M. Harries, Splice-2 comparative evaluation: Electricity pricing, Technical report, The University of South Wales, 1999. Access from J.Gama webpage. Comment on applicability .

PAKDD'09 competition data represents the credit evaluation task. It is collected over a five-year period. Unfortunately, the true labels are released only for the first part of the data. Access

Sensor stream and Power supply stream datasets are available from X. Zhu's Stream Data Mining Repository. Access

SMEAR is a benchmark data stream with a lot of missing values. Environment observation data over 7 years. Predict cloudiness. Access

Text mining , a collection of text mining datasets with concept drift, maintained by I. Katakis. Access

Gas Sensor Array Drift Dataset , a collection of 13,910 measurements from 16 chemical sensors utilized for drift compensation in a discrimination task of 6 gases at various levels of concentrations. Access

Other

KDD'99 competition data contains simulated intrusions in a military network environment. It is often used as a benchmark to evaluate handling concept drift. Access

Synthetic

Extreme verification latency benchmark Souza, V.M.A.; Silva, D.F.; Gama, J.; Batista, G.E.A.P.A. (2015). "Data Stream Classification Guided by Clustering on Nonstationary Environments and Extreme Verification Latency" . Proceedings of the 2015 SIAM International Conference on Data Mining (SDM) . SIAM. pp. 873– 881. doi : 10.1137/1.9781611974010.98 . ISBN 9781611974010 . S2CID 19198944 . Access from Nonstationary Environments – Archive.

Sine, Line, Plane, Circle and Boolean Data Sets Minku, L.L.; White, A.P.; Yao, X. (2010). "The Impact of Diversity on On-line Ensemble Learning in the Presence of Concept Drift" (PDF) . IEEE Transactions on Knowledge and Data Engineering . 22 (5): 730– 742. doi : 10.1109/TKDE.2009.156 . S2CID 16592739 . Access from L.Minku webpage.

SEA concepts Street, N.W.; Kim, Y. (2001). "A streaming ensemble algorithm (SEA) for large-scale classification" (PDF) . KDD'01: Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining . pp. 377– 382. doi : 10.1145/502512.502568 . ISBN 978-1-58113-391-2 . S2CID 11868540 . Access from J.Gama webpage.

STAGGER Schlimmer, J.C.; Granger, R.H. (1986). "Incremental Learning from Noisy Data" . Mach. Learn . 1 (3): 317– 354. doi : 10.1007/BF00116895 . S2CID 33776987 .

Mixed Gama, J.; Medas, P.; Castillo, G.; Rodrigues, P. (2004). "Learning with drift detection" . Brazilian symposium on artificial intelligence . Springer. pp. 286– 295. doi : 10.1007/978-3-540-28645-5_29 . ISBN 978-3-540-28645-5 . S2CID 2606652 .

Data generation frameworks

Minku, White & Yao 2010 Download from L.Minku webpage.

Lindstrom, P.; Delany, S.J.; MacNamee, B. (2008). "Autopilot: Simulating Changing Concepts in Real Data" (PDF) . Proceedings of the 19th Irish Conference on Artificial Intelligence & Cognitive Science . pp. 272– 263.

Narasimhamurthy, A.; Kuncheva, L.I. (2007). "A framework for generating data to simulate changing environments" . AIAP'07: Proceedings of the 25th IASTED International Multi-Conference: artificial intelligence and applications . pp. 384– 389. Code

Projects

INFER : Computational Intelligence Platform for Evolving and Robust Predictive Systems (2010–2014), Bournemouth University (UK), Evonik Industries (Germany), Research and Engineering Centre (Poland)

HaCDAIS : Handling Concept Drift in Adaptive Information Systems (2008–2012), Eindhoven University of Technology (the Netherlands)

KDUS : Knowledge Discovery from Ubiquitous Streams, INESC Porto and Laboratory of Artificial Intelligence and Decision Support (Portugal)

ADEPT : Adaptive Dynamic Ensemble Prediction Techniques, University of Manchester (UK), University of Bristol (UK)

ALADDIN : autonomous learning agents for decentralised data and information networks (2005–2010)

GAENARI : C++ incremental decision tree algorithm. it minimize concept drifting damage. (2022)

Benchmarks

NAB : The Numenta Anomaly Benchmark, benchmark for evaluating algorithms for anomaly detection in streaming, real-time applications. (2014–2018)

Meetings

2014 [] Special Session on "Concept Drift, Domain Adaptation & Learning in Dynamic Environments" @IEEE IJCNN 2014

[] Special Session on "Concept Drift, Domain Adaptation & Learning in Dynamic Environments" @IEEE IJCNN 2014

2013 RealStream Real-World Challenges for Data Stream Mining Workshop-Discussion at the ECML PKDD 2013, Prague, Czech Republic. LEAPS 2013 The 1st International Workshop on Learning stratEgies and dAta Processing in nonStationary environments

RealStream Real-World Challenges for Data Stream Mining Workshop-Discussion at the ECML PKDD 2013, Prague, Czech Republic.

LEAPS 2013 The 1st International Workshop on Learning stratEgies and dAta Processing in nonStationary environments

2011 LEE 2011 Special Session on Learning in evolving environments and its application on real-world problems at ICMLA'11 HaCDAIS 2011 The 2nd International Workshop on Handling Concept Drift in Adaptive Information Systems ICAIS 2011 Track on Incremental Learning IJCNN 2011 Special Session on Concept Drift and Learning Dynamic Environments CIDUE 2011 Symposium on Computational Intelligence in Dynamic and Uncertain Environments

LEE 2011 Special Session on Learning in evolving environments and its application on real-world problems at ICMLA'11

HaCDAIS 2011 The 2nd International Workshop on Handling Concept Drift in Adaptive Information Systems

ICAIS 2011 Track on Incremental Learning

IJCNN 2011 Special Session on Concept Drift and Learning Dynamic Environments

CIDUE 2011 Symposium on Computational Intelligence in Dynamic and Uncertain Environments

2010 HaCDAIS 2010 International Workshop on Handling Concept Drift in Adaptive Information Systems: Importance, Challenges and Solutions ICMLA10 Special Session on Dynamic learning in non-stationary environments SAC 2010 Data Streams Track at ACM Symposium on Applied

Computing SensorKDD 2010 International Workshop on Knowledge Discovery from Sensor Data StreamKDD 2010 Novel Data Stream Pattern Mining Techniques Concept Drift and Learning in Nonstationary Environments at IEEE World Congress on Computational Intelligence MLMDS’2010 Special Session on Machine Learning Methods for Data Streams at the 10th International Conference on Intelligent Design and Applications, ISDA’10

HaCDAIS 2010 International Workshop on Handling Concept Drift in Adaptive Information Systems: Importance, Challenges and Solutions

ICMLA10 Special Session on Dynamic learning in non-stationary environments

SAC 2010 Data Streams Track at ACM Symposium on Applied Computing

SensorKDD 2010 International Workshop on Knowledge Discovery from Sensor Data

StreamKDD 2010 Novel Data Stream Pattern Mining Techniques

Concept Drift and Learning in Nonstationary Environments at IEEE World Congress on Computational Intelligence

MLMDS’2010 Special Session on Machine Learning Methods for Data Streams at the 10th International Conference on Intelligent Design and Applications, ISDA’10

References