International Journal of Grid and Distributed Computing
Vol. 9, No. 4 (2016), pp.135-150
http://dx.doi.org/10.14257/ijgdc.2016.9.4.13

Link Prediction for Authorship Association in Heterogeneous Network Using Streaming Classification

Harshal Singh, Divya Tomar and Sonali Agarwal

Indian Institute of Information Technology, Allahabad, India
ise2013012@iiita.ac.in , divyatomar26@gmail.com and sonali@iiita.ac.in

Abstract

Prediction of links or relations between the objects in any network is no longer a new task these days; in fact it has become a high rated area of research and has attracted many researchers seeking their contribution to the mentioned area. Research has seen an exponential growth over the passing years, and the active researchers do not hesitate in linking with fellow researchers working in same domain irrespective of their geographic location. However this in turn has generated a very complex network of objects and links which are needed to be analyzed and dealt with. Prediction of co-authorship is the sub domain of link prediction and with the increasing complexity of co-authorship network the authors are treated as heterogeneous entity not as homogeneous ones. The rule is simple analyze the data preprocess it, train the classifier according to desired classification rules and then get the classified form of data. But irrelevant features always reflect various impacts and issues on generation of a classifier and consequently the impact is sustained to further classification results. Therefore, this paper proposes streaming classification algorithm combined with Correlation based Feature selection as a solution to the stated problem. The consistent and relevant features are selected with the help of feature selection algorithm and then these features are classified with the help of streaming classification algorithm- Very Fast Decision Tree (VFDT). VFDT is a streaming classification algorithm and it takes the dataset in the form of continuous stream as an input. Finally the effectiveness of the proposed algorithm can be seen in the experimental results.

Keywords: *Very Fast Decision Tree, Link Prediction, Streaming Classification*

1. Introduction

This research work focuses on co-authorship prediction which means binding authors together so that best quality research can be promoted with proper team effort [1-2]. Co-authorship network is basically formed when various authors interact and collaborate with each other thus forming a complete co-authorship network. It is an evident fact that researches that is carried out by potential researchers produce better and more fruitful results when taken together as compared to the research being carried out by individual researcher. Many times, it is found that researcher wants to collaborate with other researchers of same domain and interest. The major aspect of link prediction for co-authorship network is to predict how the same network will evolve [1]. This will help in several ways; (i) giving an insight of the structure of how the scientific collaborations are taking place; (ii) finding out various research and related communities and how they are evolving, this will further help in promoting quality and focused research; (iii) exploring the potential research areas of scientists so that actual doers get benefited; and (iv) it also helps individual researchers in finding their colleagues, companions, students or assistants.

The diagrammatic representation of link prediction is shown in Figure 1. This is a snapshot of a social network at various intervals of time starting from t_i to t_k. The edges

ISSN: 2005-4262 IJGDC

are the links between various nodes, as we can see how the network is evolving and how the nodes are getting linked, now we need to predict what will be the state of network at time t_k, what new links will be added and how the network further changes. The domain of the research work is co-authorship prediction, for that just consider the nodes as authors and there is a need to find the various links among the authors that is who are the authors those are going to be linked.

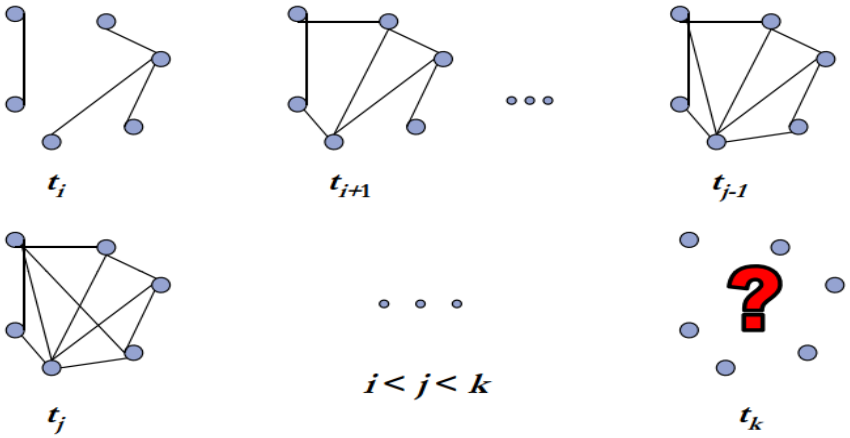


Figure 1. Representation of Link Prediction

The problem of link prediction is one of the classic examples of machine learning. There are numerous domains of link prediction for example link prediction for social network [1], link prediction for medical network [3] *etc*. There are various parameters on the basis of which researchers can be associated with each other but finding researchers which can be associated with each other whether in the same department of particular organization or not is not always easy. Presently, the academic collaborative network is getting bigger and complex as more and more researchers, teachers, and academicians are participating in academic communication. With ever increasing tools, techniques and other resources, it is very much possible that research is going to expand. So the best possible way to contribute in the new rising scenario is to lay down a road for better interaction and further improved collaboration. Majority of the techniques proposed for the co-authorship prediction treats the network as homogeneous [1-7] and takes the data for input as static data but it is an evident fact that the exhaustive and intense research makes the complete academic network heterogeneous and dynamic because it is changing and increasing with time. The current scenario has brought up a huge change nowadays. The architecture is changing and increasing rapidly therefore there is a need an overall solution when it comes to finding co-authors, so that quality and new innovations in the field of research can be produced and can lead to substantial development. Therefore this paper proposes a complete approach of dynamic classification with stream learning classifier Very Fast Decision Tree [8-10] for heterogeneous network and to further improve the classification results the correlation based feature selection [11] algorithm is used. Feature selection algorithm has some significant advantages like removing irrelevant and inconsistent features. Irrelevant and inconsistent features are those which do not significantly add up to result or accuracy and hence they can be removed so that classification task can be improved. VFDT is a dynamic data classification algorithm and has various significant advantages such as it has a higher data processing rate and can possibly handle infinite amount of data. As compared to the traditional classification algorithms, learning probability of the VFDT is high and learning task is done much faster because VFDT learns from data streams instead of database. In addition to above stated

advantages VFDT produces significant results when compared to various other classification algorithms such as Decision Tree, Neural Network and k-Nearest Neighbor.

Rest of the paper is organized as follows. Section 2 discusses the literature survey in the field of link prediction. Section 3 provides the detailed description of the proposed methodology. Experimental results are discussed in section 4 and finally conclusion and future recommendation is given in section 5.

2. Literature Survey

This research work divided the literature survey into two categories:

- Literature survey of the problem domain
- Literature survey of the various link prediction techniques.

2.1 Literature Survey of the Problem Domain

The problem of link prediction has always been a debatable topic and it has many application areas also. Some of the applications are discussed here:

2.1.1 Identification of the Structure of a Criminal Network: This one is a serious issue as it has a greater social impact. With the help of the incomplete data the missing links can be predicted and it can be of great help to identify the structure of a complete criminal network [12]. This will help in finding the criminals because various criminals, on the basis of their various properties, will be linked together thereby helping the agencies to find and fetch them.

2.1.2 Developing a Recommender System for the Customers: A proper recommender system [13] can help any business organization in increasing its sales and hence profit. It has a high commercial impact. Today almost every e-commerce site uses a properly developed and feature rich recommender system. A recommender system suggests the customer with most predictable next options that can be looked into or that can also be purchased with the article of their choice. But how does it suggest something related. This can be achieved with the help of link prediction. All the entities in a store are considered as nodes and various possible links are found between them according to the topological features of the products. Based on these features all the linked products are suggested to the customers by various sites when they land on their home pages or during purchasing. This helps both the parties involved in business one who is buying and the other for whom one is buying.

2.1.3 Predictive Server Pre-Fetching: Internet has grown exponentially since its advent. Various applications such e-commerce and social media has further alleviated the ever rising scenario. Server pre fetching has an integral use in e commerce. Any website developer and e-commerce manager has two main concerns while developing any ecommerce website: content that is to be put on the website and the performance of the website. The content mostly holds the information about the products and the services offered by that website, and the content holds the customer. So very careful thought process should be applied for the content that is to be placed on the website whereas the performance (in terms of response time) has to be great, because if the site takes much time to load or to process customer request will irritate the customer and the company will lose the potential customer. Pre fetching [14] can be an answer to this situation. If it can be known, what documents user will request from the web then these documents can be pre loaded in cache to improve the latency and this can be done with the help of link prediction.

2.1.4 Accelerating Academic Collaborative Network: In this research work, the application that is being dealt with, is that how academic collaboration or association can be improved and facilitated [15]. Link prediction finds out various common topological features and predicts the association that can be possible among various academicians.

2.2 Literature Survey of Various Link Prediction Techniques: There are numerous techniques for the prediction of links [7] but they are categorized into three main strategies- similarity based strategy, maximum likelihood strategy, and probabilistic strategy.

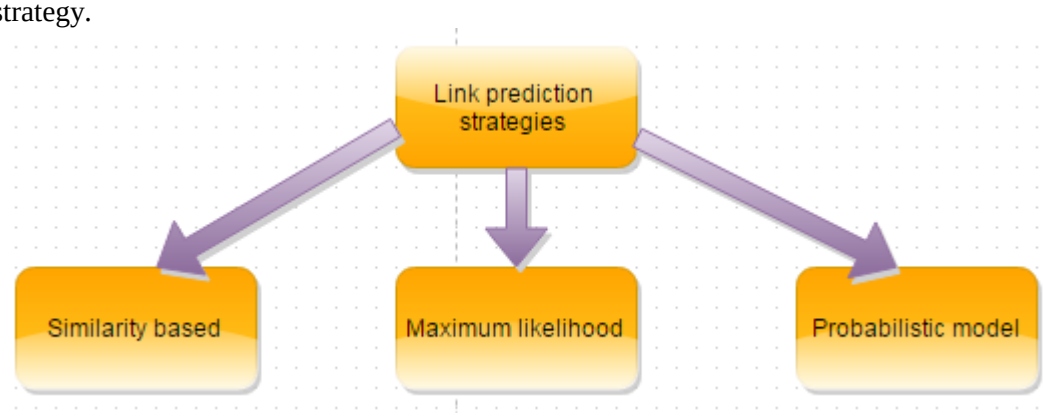


Figure 2. Block Diagram for Link Prediction Strategies

2.2.1 Similarity based Strategies: The simplest approach of all the approaches of link prediction is the similarity-based algorithm [1]. In this method every pair of node has a score associated to it. It can be understood by assigning the score S_{ab} to the nodes 'a' and 'b'. This score is nothing but the measure of similarity or proximity between 'a' and 'b'. Likewise all the pair of nodes or every one of the connections are doled out the score and afterward they are positioned by scores. The connections which associate more comparable nodes are considered to have higher presence of probability. The similarity index can be easily applied to some topological network but may fail with various other networks. The actual working of similarity based algorithms is to find out the similarity between nodes and consequently links are generated between similar nodes. The similarity between two nodes is determine by finding common features or attributes. Finding attribute similarity is not an easy task because the attributes of nodes are generally not known, they are hidden. Therefore the area of concern is finding structural similarity which is governed by the structure of the network. There are various classification levels of structural similarity indices such as: parameter dependent vs. parameter independent, local vs. global, node dependent vs. edge dependent *etc*. The other classification level for consideration is structural equivalence and regular equivalence. In case of structural equivalence, it follows an assumption that similarity between two end points is indicated by link itself. In case of regular equivalence, it is assumed that if the neighbors of two nodes are similar then the nodes are also considered similar.

2.2.2 Maximum Likelihood Methods [16]: The working of maximum likelihood methods can be understood as: the user predefines some organizing principles and rules according to the structure of the given network. The organizing principles mainly describe detailed rules and all the essential parameters. The organizing principles for present nodes are calculated by maximizing the likelihood of the given network structure. Then all the non connected links are also observed and the likelihood for these links is calculated based on the previously obtained organizing principles. The main drawbacks of this structure are that, first of all the maximum likelihood methods are very time consuming

because obtaining organizing principles is not an easy and fast task, which makes these methods time intensive. Other drawback being theses methods can handle small networks with few thousand nodes. Even well designed algorithms can handle small networks in reasonable amount of time and generally fails when it comes to handling huge networks probably containing millions of nodes. The accuracies obtained by the maximum likelihood algorithms are also not good and also not comparable to various other existing techniques of link prediction. Their advantage is that they provide better understanding and insight of the network structure, as compared to the probabilistic methods and similarity-based methods. There are two recently proposed algorithms that are needed to be studied under this category.

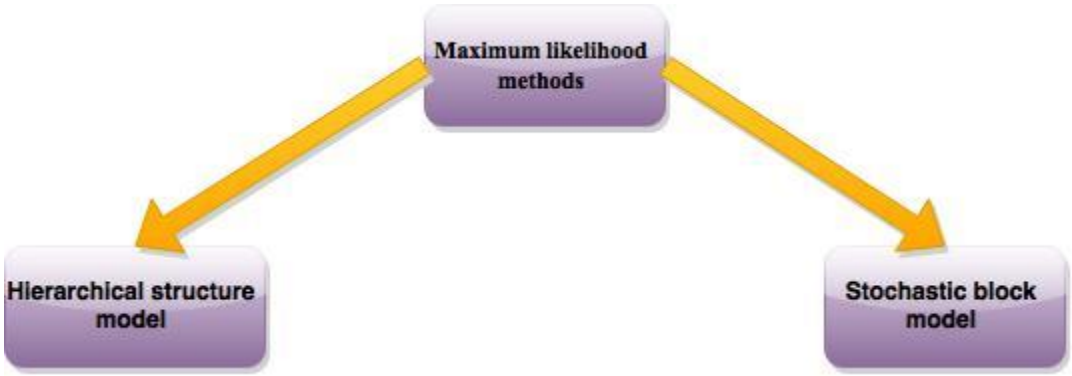


Figure 3. Block Diagram for Maximum Likelihood Methods

- **Hierarchical Structure Model**

  There are numerous evidences which show that various real network structures follow a hierarchical pattern or structure. The nodes are divided into sub nodes and then these sub nodes are further divided into other sub nodes thus forming a hierarchy of nodes [17]. The hierarchical structure follows a division to multiple scales.

- **Stochastic block model**

  Stochastic block model is another classification of maximum likelihood methods [18-24]. This method holds its name for its generic characteristics, and is the most generic model for link prediction. The stochastic block model works by partitioning the nodes into various groups and subgroups. The nodes with similar likelihood scores are supposed to fall into same groups and the probability that two nodes get connected or linked depends totally on the groups or sub groups under which these nodes fall. The main advantage of the stochastic block model can be seen in the community center [25], connections which are role to role [26-27], or air transportation network where the splitting of nodes can be easily done into various groups and sub groups. This splitting can be done because of the generic characterstics and thus the linking of various other non connected and non linked nodes can also be done to the groups where these nodes belong.

2.2.3 Probabilistic Model or Supervised Learning Model

Probabilistic model or supervised learning model can be understood with relation to human learning [28]. Human gain from its past experience however there is in no way like past encounters in the event of computer adapting, so the computer learns with the assistance of information. There are two angles in administered learning preparing and testing. The computer learns from the set of input data or training data and a vector is inferred based on the training data. This vector is known as a classifier and based on the input data various rules are defined for the classification of upcoming data. The input data

to be studied unfolds various aspects about the structure of the network. Some of the knowledge that is gained can have some or all of the following: the path length, knowledge about the neighbors, and number of links that can be formed between various nodes. With the good learning process the inferred vector that is the classifier if trained and then it has the capability to classify any set of input data based on the rules formed during training time. The efficiency of any classifier is measured in terms of the accuracy.

3. Proposed Methodology

This research work proposes streaming classification algorithm combined with correlation based feature selection as a solution to the link prediction for co-authorship network. The consistent and relevant features are selected with the help of feature selection algorithm and then these features are classified with the help of streaming classification algorithm-Very Fast Decision Tree. The classification algorithm is a streaming classification algorithm and it takes the dataset in the form of continuous stream as an input. Figure 4 shows the block diagram of proposed methodology.

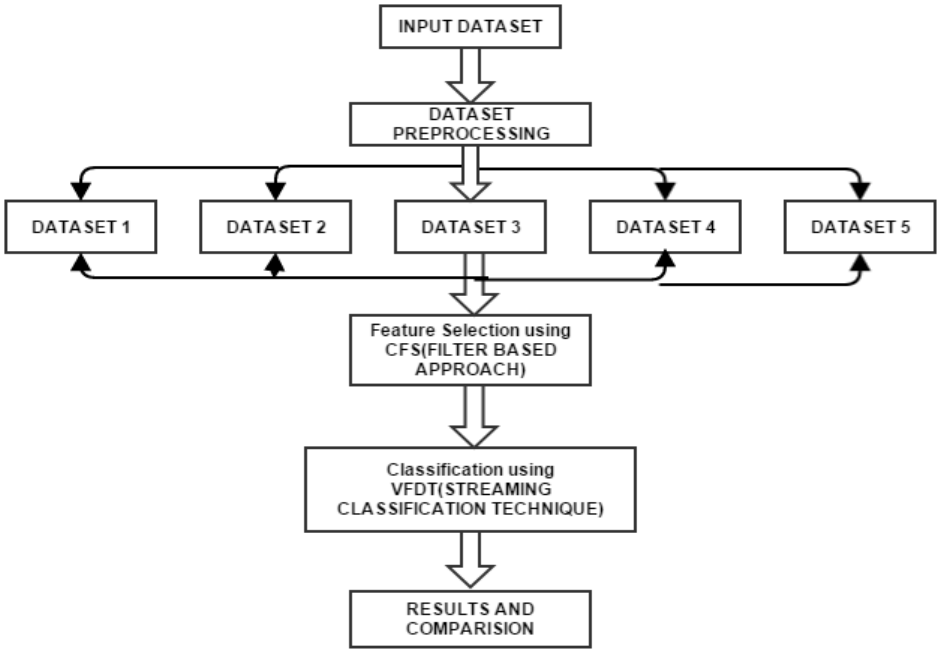


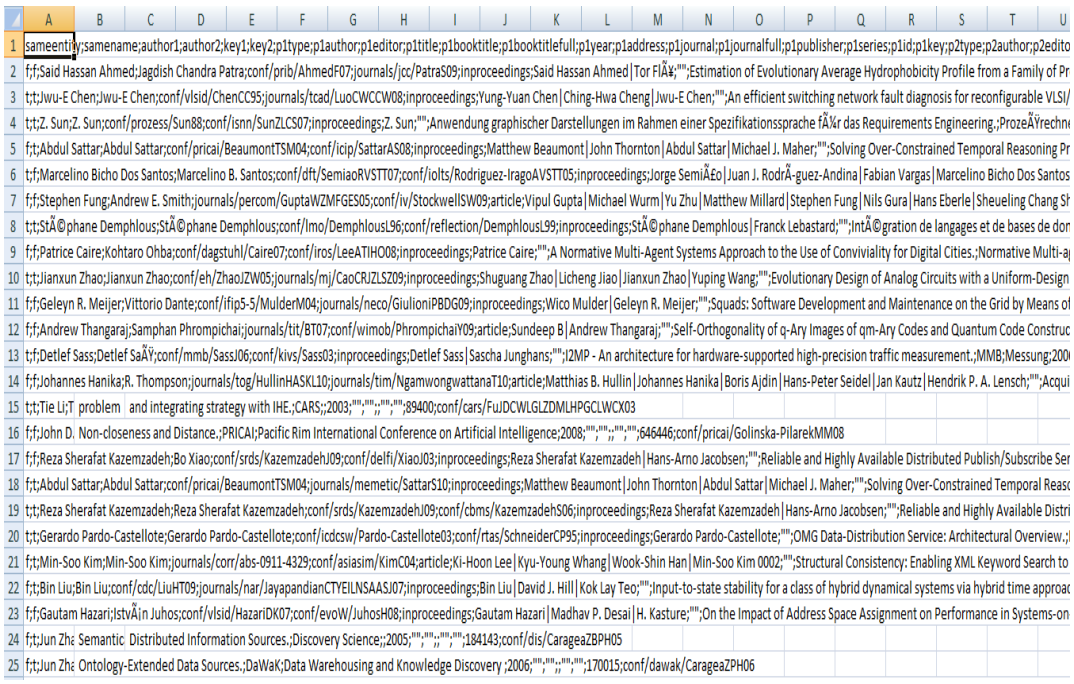
Figure 4. Block Diagram for the Proposed Work

3.1 Dataset

The dataset which is used for experimentation is the DBLP dataset. The dataset gives sufficient information about the proceedings and computer science journals. The complete downloaded dataset has 34 columns each representing particular attribute. The dataset can be downloaded from: http://hpi.de/en/naumann/projects/repeatability/datasets/dblp-dataset.html This is how the actual dataset looks. Figure 5 shows the snapshot of the dataset before pre-processing. This dataset at this point of time is unprocessed and requires preprocessing. The detailed description of the dataset is given below:

- **sameentity(boolean):** Here author1 is represented as same entity as the author2.
- **samename(boolean):** It shows that name of author1 and author2 is same.
- **authorname1, authorname2 (string):** Names of the authors which are to be compared.

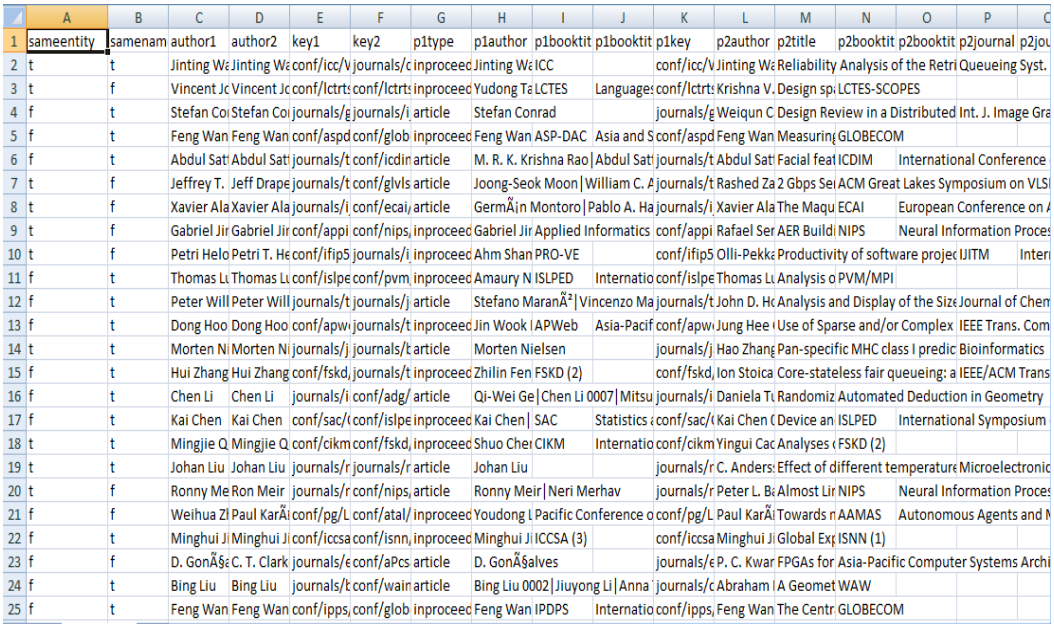
- **p1*, p2*(string):** It represents the details of the publications(p1,p2) that were compared in the DBLP database.
- **p[1/2]booktitlefull, p[1/2]journalfull(string):** It represents the complete names of journals and books titles abbreviations contained in the DBLP dataset.
- **p[1/2][author/editor](string):** These are the values for multi-valued attributes which are authors and editors. These values are separated by pipe symbol.



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
1	sameentity;samename;author1;author2;key1;key2;p1type;p1author;p1editor;p1title;p1booktitle;p1booktitlefull;p1year;p1address;p1journal;p1journalfull;p1publisher;p1series;p1id;p1key;p2type;p2author;p2edito																				
2	f;f;Said Hassan Ahmed;Jagdish Chandra Patra;conf/prib/AhmedF07;journals/jcc/PatraS09;inproceedings;Said Hassan Ahmed|Tor FlÃ¥;"";Estimation of Evolutionary Average Hydrophobicity Profile from a Family of Pr																				
3	t;t;Jwu-E Chen;Jwu-E Chen;conf/vlsid/ChenCC95;journals/tcad/LuoCWCCW08;inproceedings;Yung-Yuan Chen|Ching-Hwa Cheng|Jwu-E Chen;"";An efficient switching network fault diagnosis for reconfigurable VLSI/																				
4	t;t;Z. Sun;Z. Sun;conf/prozess/Sun88;conf/isnn/SunZLCS07;inproceedings;Z. Sun;"";Anwendung graphischer Darstellungen im Rahmen einer Spezifikationssprache fÃ¼r das Requirements Engineering.;ProzeÃŸrechn																				
5	f;t;Abdul Sattar;Abdul Sattar;conf/pricai/BeaumontTSM04;conf/icip/SattarAS08;inproceedings;Matthew Beaumont|John Thornton|Abdul Sattar|Michael J. Maher;"";Solving Over-Constrained Temporal Reasoning Pr																				
6	t;f;Marcelino Bicho Dos Santos;Marcelino B. Santos;conf/dft/SemiaoRVSTT07;conf/iolts/Rodriguez-IragoAVSTT05;inproceedings;Jorge SemiÃ£o|Juan J. RodrÃ-guez-Andina|Fabian Vargas|Marcelino Bicho Dos Santos																				
7	f;f;Stephen Fung;Andrew E. Smith;journals/percom/GuptaWZMFGES05;conf/iv/StockwellSW09;article;Vipul Gupta|Michael Wurm|Yu Zhu|Matthew Millard|Stephen Fung|Nils Gura|Hans Eberle|Sheueling Chang Sh																				
8	t;t;StÃ©phane Demphlous;StÃ©phane Demphlous;conf/lmo/DemphlousL96;conf/reflection/DemphlousL99;inproceedings;StÃ©phane Demphlous|Franck Lebastard;"";Intégration de langages et de bases de don																				
9	f;f;Patrice Caire;Kohtaro Ohba;conf/dagstuhl/Caire07;conf/iros/LeeATIHO08;inproceedings;Patrice Caire;"";A Normative Multi-Agent Systems Approach to the Use of Conviviality for Digital Cities.;Normative Multi-a																				
10	t;t;Jianxun Zhao;Jianxun Zhao;conf/eh/ZhaoJZW05;journals/mj/CaoCRJZLSZ09;inproceedings;Shuguang Zhao|Licheng Jiao|Jianxun Zhao|Yuping Wang;"";Evolutionary Design of Analog Circuits with a Uniform-Design																				
11	f;f;Geleyn R. Meijer;Vittorio Dante;conf/ifip5-5/MulderM04;journals/neco/GiulioniPBDG09;inproceedings;Wico Mulder|Geleyn R. Meijer;"";Squads: Software Development and Maintenance on the Grid by Means of																				
12	f;f;Andrew Thangaraj;Samphan Phrompichai;journals/tit/BT07;conf/wimob/PhrompichaiY09;article;Sundeep B|Andrew Thangaraj;"";Self-Orthogonality of q-Ary Images of qm-Ary Codes and Quantum Code Construc																				
13	t;f;Detlef Sass;Detlef SaÃŸ;conf/mmb/SassJ06;conf/kivs/Sass03;inproceedings;Detlef Sass|Sascha Junghans;"";J2MP - An architecture for hardware-supported high-precision traffic measurement.;MMB;Messung;200																				
14	f;f;Johannes Hanika;R. Thompson;journals/tog/HullinHASKL10;journals/tim/NgamwongwattanaT10;article;Matthias B. Hullin|Johannes Hanika|Boris Ajdin|Hans-Peter Seidel|Jan Kautz|Hendrik P. A. Lensch;"";Acqui																				
15	t;t;Tie Li;T	problem	and integrating strategy with IHE.;CARS;;2003;"";"";;;"";"";89400;conf/cars/FuJDCWLGLZDMLHPGCLWCX03																		
16	f;f;John D.	Non-closeness and Distance.;PRICAI;Pacific Rim International Conference on Artificial Intelligence;2008;"";"";;"";"";"";646446;conf/pricai/Golinska-PilarekMM08																			
17	f;f;Reza Sherafat Kazemzadeh;Bo Xiao;conf/srds/KazemzadehJ09;conf/delfi/XiaoJ03;inproceedings;Reza Sherafat Kazemzadeh|Hans-Arno Jacobsen;"";Reliable and Highly Available Distributed Publish/Subscribe Ser																				
18	f;t;Abdul Sattar;Abdul Sattar;conf/pricai/BeaumontTSM04;journals/memetic/SattarS10;inproceedings;Matthew Beaumont|John Thornton|Abdul Sattar|Michael J. Maher;"";Solving Over-Constrained Temporal Reasc																				
19	t;t;Reza Sherafat Kazemzadeh;Reza Sherafat Kazemzadeh;conf/srds/KazemzadehJ09;conf/cbms/KazemzadehS06;inproceedings;Reza Sherafat Kazemzadeh|Hans-Arno Jacobsen;"";Reliable and Highly Available Distri																				
20	t;t;Gerardo Pardo-Castellote;Gerardo Pardo-Castellote;conf/icdcsw/Pardo-Castellote03;conf/rtas/SchneiderCP95;inproceedings;Gerardo Pardo-Castellote;"";OMG Data-Distribution Service: Architectural Overview.;I																				
21	f;t;Min-Soo Kim;Min-Soo Kim;journals/corr/abs-0911-4329;conf/asiasim/KimC04;article;Ki-Hoon Lee|Kyu-Young Whang|Wook-Shin Han|Min-Soo Kim 0002;"";Structural Consistency: Enabling XML Keyword Search to																				
22	f;t;Bin Liu;Bin Liu;conf/cdc/LiuHT09;journals/nar/JayapandianCTYEILNSAASJ07;inproceedings;Bin Liu|David J. Hill|Kok Lay Teo;"";Input-to-state stability for a class of hybrid dynamical systems via hybrid time approac																				
23	f;f;Gautam Hazari;IstvÃ¡n Juhos;conf/vlsid/HazariDK07;conf/evoW/JuhosH08;inproceedings;Gautam Hazari|Madhav P. Desai|H. Kasture;"";On the Impact of Address Space Assignment on Performance in Systems-on																				
24	f;t;Jun Zha	Semantic	Distributed Information Sources.;Discovery Science;;2005;"";"";;"";"";184143;conf/dis/CarageaZBPH05																				
25	f;t;Jun Zha	Ontology-Extended Data Sources.;DaWaK;Data Warehousing and Knowledge Discovery ;2006;"";"";;"";"";170015;conf/dawak/CarageaZPH06																				

Figure 5. Dataset before Pre Processing

This is the description of the various columns of the dataset. The dataset after preprocessing task will become the first input to the task undertaken. The preprocessing task mainly included manual preprocessing; there were many missing values in the dataset which are handled manually. The dataset after preprocessing is shown in Figure 6 as follows:



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	C
1	sameentity	samenam	author1	author2	key1	key2	p1type	p1author	p1booktit	p1booktit	p1key	p2author	p2title	p2booktit	p2booktit	p2journal	p2jou
2	t	t	Jinting Wa	Jinting Wa	conf/icc/V	journals/c	inproceed	Jinting Wa	ICC		conf/icc/V	Jinting Wa	Reliability	Analysis of the Retri		Queueing Syst.	
3	t	f	Vincent Jc	Vincent Jc	conf/lctrts	conf/lctrts	inproceed	Yudong Ta	LCTES	Languages	conf/lctrts	Krishna V.	Design sp	LCTES-SCOPES			
4	f	t	Stefan Co	Stefan Co	journals/g	journals/i	article	Stefan Conrad			journals/g	Weiqun C	Design Review in a Distributed			Int. J. Image Gra	
5	f	t	Feng Wan	Feng Wan	conf/aspd	conf/glob	inproceed	Feng Wan	ASP-DAC	Asia and S	conf/aspd	Feng Wan	Measurin	GLOBECOM			
6	f	t	Abdul Sat	Abdul Sat	journals/t	conf/icdin	article	M. R. K. Krishna Rao	Abdul Sat		journals/t	Abdul Sat	Facial feat	ICDIM	International Conference		
7	t	f	Jeffrey T.	Jeff Drape	journals/t	conf/glvls	article	Joong-Seok Moon	William C. A		journals/t	Rashed Za	2 Gbps Se	ACM Great Lakes Symposium on VLS			
8	t	f	Xavier Ala	Xavier Ala	journals/i	conf/ecai/	article	GermÃ¡n Montoro	Pablo A. Ha		journals/i	Xavier Ala	The Maqu	ECAI	European Conference on A		
9	t	f	Gabriel Jir	Gabriel Jir	conf/appi	conf/nips/	inproceed	Gabriel Jir	Applied Informatics		conf/appi	Rafael Ser	AER Build	NIPS	Neural Information Proces		
10	t	f	Petri Helo	Petri T. He	conf/ifip5	journals/i	inproceed	Ahm Shan	PRO-VE		conf/ifip5	Olli-Pekka	Productivity of software projec			IJITM	Inter
11	f	t	Thomas Lu	Thomas Lu	conf/islpe	conf/pvm/	inproceed	Amaury N	ISLPED	Internatio	conf/islpe	Thomas Lu	Analysis o	PVM/MPI			
12	f	t	Peter Will	Peter Will	journals/t	journals/j	article	Stefano MaranÃ²	Vincenzo Ma		journals/t	John D. Hc	Analysis and Display of the Size			Journal of Chem	
13	f	t	Dong Hoo	Dong Hoo	conf/apw	journals/t	inproceed	Jin Wook	APWeb	Asia-Pacif	conf/apw	Jung Hee	Use of Sparse and/or Complex			IEEE Trans. Com	
14	t	t	Morten Ni	Morten Ni	journals/j	journals/k	article	Morten Nielsen			journals/j	Hao Zhang	Pan-specific MHC class I predic			Bioinformatics	
15	f	t	Hui Zhang	Hui Zhang	conf/fskd,	journals/t	inproceed	Zhilin Fen	FSKD (2)		conf/fskd,	Ion Stoica	Core-stateless fair queueing: a			IEEE/ACM Trans	
16	f	t	Chen Li	Chen Li	journals/i	conf/adg/	article	Qi-Wei Ge	Chen Li 0007	Mitsu	journals/i	Daniela Tu	Randomiz	Automated Deduction in Geometry			
17	f	t	Kai Chen	Kai Chen	conf/sac/(	conf/islpe	inproceed	Kai Chen	SAC	Statistics a	conf/sac/(	Kai Chen (	Device an	ISLPED	International Symposium		
18	t	t	Mingjie Q	Mingjie Q	conf/cikm	conf/fskd,	inproceed	Shuo Chen	CIKM	Internatio	conf/cikm	Yingui Cac	Analyses (	FSKD (2)			
19	t	t	Johan Liu	Johan Liu	journals/r	journals/r	article	Johan Liu			journals/r	C. Anders	Effect of different temperature			Microelectronic	
20	t	f	Ronny Me	Ron Meir	journals/r	conf/nips/	article	Ronny Meir	Neri Merhav		journals/r	Peter L. Ba	Almost Lir	NIPS	Neural Information Proces		
21	f	f	Weihua Zl	Paul KarÃ	conf/pg/L	conf/atal/	inproceed	Youdong L	Pacific Conference o		conf/pg/L	Paul KarÃ	Towards n	AAMAS	Autonomous Agents and M		
22	f	t	Minghui Ji	Minghui Ji	conf/iccsa	conf/isnn/	inproceed	Minghui Ji	ICCSA (3)		conf/iccsa	Minghui Ji	Global Exp	ISNN (1)			
23	f	f	D. GonÃ§a	C. T. Clark	journals/e	conf/aPcs	article	D. GonÃ§alves			journals/e	P. C. Kwar	FPGAs for	Asia-Pacific Computer Systems Archi			
24	f	t	Bing Liu	Bing Liu	journals/t	conf/wain	article	Bing Liu 0002	Jiuyong Li	Anna	journals/c	Abraham I	A Geomet	WAW			
25	f	t	Feng Wan	Feng Wan	conf/ipps/	conf/glob	inproceed	Feng Wan	IPDPS	Internatio	conf/ipps/	Feng Wan	The Centr	GLOBECOM			

Figure 6. Dataset after Pre Processing

3.2 Feature Selection using CFS Algorithm

Feature selection is mainly carried out to remove irrelevant and redundant features (which do not provide any additional information than the selected attributes) from the input dataset so that classification accuracy can be improved and classification time is also reduced because feature selection algorithm reduces the degree of the input dataset [29-37]. The feature selection algorithms are basically classified into two broader types, wrapper based feature selection approach and filter based feature selection approach [38]. Both the techniques reduces the degree of input but their working approach is different, wrapper based algorithms finds out the importance of features by using learning based approach [39-40]. A predictive model is used for scoring feature subsets which in turn trains a new model for every subset thus making wrapper based approach very computationally intensive because every time a new model is trained for every learning and classification algorithm, the results produced by wrapper based algorithms are also not generic therefore making the algorithm less susceptible to scaling for large databases. Whereas the filter based approach finds out the worth of features with the help of heuristics and general features or characteristics of the input data. There are many advantages of filter based approach over wrapper based approach thus making us convinced to go for filter based approach rather than wrapper based. Filter based approach are generic in nature and feature are also loosely coupled thus there is no need to re-execute the feature selection algorithm for every learning algorithm, this makes filter based algorithms works faster as compared to wrappers and hence they can be used for scaling to large databases. Correlation based feature selection(CFS) falls under the category of filter based approach for feature selection and there is no need for the user to pre specify any threshold or the number of attributes to be selected everything is taken care of because CFS is an automated algorithm [31, 37]. There are some other advantages of using CFS algorithm for feature selection such as it works on the complete and original feature set thus allowing the learning algorithm to interpret knowledge in terms of original features rather than in terms of transformed or modified space. It incurs very less computational costs as compared to most of the wrapper feature selection approaches. CFS works on the principle of subset features being highly correlated with class and

uncorrelated with each other. CFS like various other feature selection algorithm follows four main steps to perform the feature selection task as shown in Figure 7:

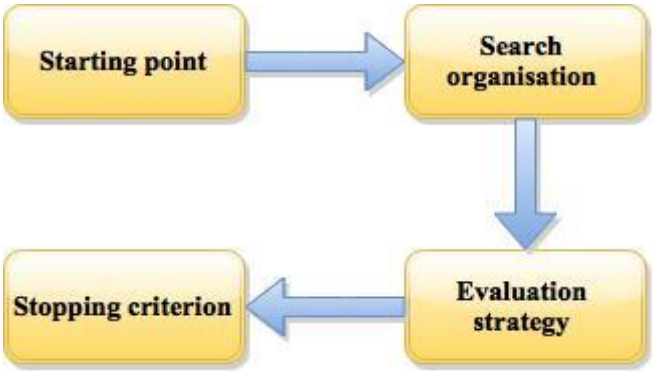


Figure 7. Architecture of CFS

CFS considers the correlation of feature with the target feature and selects only those features which show a strong correlation with the target feature and weak correlation with each other. Correlation can be estimated from the training data samples as

$$r_{zc} = \frac{kr_{zi}}{\sqrt{(k+k(k-1)r_{ii})}} \quad (1)$$

Here, r_{zc} represent the correlation between the outside variable and summed components. k represents the total number of available components. r_{zi} represents the average of correlations between the outside variables and various component. r_{ii} is the average of the inter correlation between various components. The inference that can be drawn from the above formula regarding CFS is:

- Greater the value of correlation between outside variables and various components; greater will be the correlation between outside variable and composite variable.
- Lower the value of average of the inter correlation between various components; greater will be the correlation between outside variable and composite variable.

After the application of CFS algorithm the dataset is reduced, and only the significant and relevant attributes gets selected which will contribute in the machine learning and classification task.

3.3 Very Fast Decision Tree (VFDT) Algorithm

Hoeffding tree is used as acronym for Very Fast Decision Tree though they differ slightly [8,10]. VFDT is known as the pioneer in taking care of the information in constant or in type of streams. Since the issue of fast of stream information, limitless measure of information, instability of impending stream settles on decision tree unsatisfactory for streaming data. Henceforth, we need to pick an arrangement that can deal with these issues identified with streaming data. Hoeffding tree holds its significant part in the matter of taking care of dynamic and gigantic information which is unrealistic with other order calculations which started to before Hoeffding tree idea. Hoeffding tree can basically handle vast measure of information expending even less memory at this very moment and produces comes about speedier than numerous other characterization calculations. Hoeffding trees are being mulled over in light of the fact that they speak to present cutting edge for ordering rapid streaming data. The calculation satisfies the prerequisites fundamental for adapting to streaming data while staying proficient, an accomplishment that was uncommon before its presentation. Hoeffding tree is likewise one of the incremental learning systems utilized for prediction purposes as a part of which

we have streaming data that comes uncertainly and incrementally out classifier gains from the new preparing information sets accessible. Indeed, even the best calculations today, concentrate on mining the expansive databases with the goal that they can be gained in the primary memory of the framework. Consider the sample of ATM, recording exchanges, or the telecom organizations interfacing clients by their calls and different sites that get a large number of hit day by day, and the primary thought emerges when we come to realize that the rate of increment of information is likewise expanding exponentially. The most effective frameworks grew today can just handle the issue of present situation. Hoeffding tree has a property of learning for every sample and that too in consistent time though any ordinary classifier would require numerous illustrations to create same level of learning. On the other hand, VFDT takes less time to mine the cases than the time taken by it to enter those illustrations from circle. By seeing each illustration just once, the VFDT can learn and classify. This property of VFDT additionally makes it information/yield bound. VFDT gets to be at whatever time and prepared to utilize calculation soon after seeing and gaining from beginning couple of cases.

Every internal node of a standard decision tree contains a test to isolate the cases, sending illustrations down diverse ways relying upon the estimations of specific characteristics. The pivotal choice expected to develop a decision tree is when to part a node. There is a renowned and established criterion of deciding when and where to split the decision tree. The information gain calculated for every subset of a split helps in calculating the mean of "purity". This purity of subsets is measured utilizing entropy, which for a dissemination of the class labels comprising of fractions$P_1, P_2, P_3,\ldots P_n$ adding to 1, is figured thus:

$$Entropy(P_1, P_2, \ldots, P_n) = -\sum pi \log_2 pi \quad (2)$$

The information gain is calculated by having the difference of the weighted normal entropy of the subsets of a split and the entropy of the class conveyance before the splitting. Entropy is an idea from data hypothesis that measures the amount of data passed on by a message in bits. The most effective method to make the same (or very much alike) choice in the information stream setting is the development of Hoeffding bound, also called an additive Chernoff bound. This bound is valuable on the grounds that it remains constant paying little heed to the dissemination producing the qualities, and depends just on the scope of qualities, number of perceptions and wanted certainty. A hindrance of being so broad is that it is more traditionalist as compared to distribution-dependent bound. The Hoeffding bound can be illustrated with the help of the given formula:

$$\varepsilon = \sqrt{\frac{R^2 \ln(1/\delta)}{2n}} \quad (3)$$

Here, ε is referred as Hoeffding bound. R is known as a random variable, which in this case has range log(c) and c defines the actual number of the labels of classes given in the dataset used. 1-δ defines the probability of confidence. $\bar{r}$ - ε is referred as the true/actual mean of variable. Some critical parameters and facts in reference to Hoeffding tree are listed as follows:

- G (x_i) is defined as the heuristic measure for choosing the test attribute for the present leaf node.
- G (x_i) could be the Gini index or it can be the Information Gain.
- The node with the highest value of G can act as the attribute which governs the criteria for splitting the present leaf node.
- Here G(x_a) is the attribute having highest value of the heuristic measure and G(x_b) is the attribute having second highest value of the heuristic measure

- D = G (x_a) – G (x_b) represents the difference of the first highest and the second highest attribute.
- On the off chance that the estimation of D comes out to be greater as compared to the value of the Hoeffding bound then we can easily split on the account of attributex_a.
- ΔG = G(x_a) – G(x_b)
- If ΔG > € then we can say that Hoeffding bound has chosen x_a as the correct attribute for splitting.
- Since the equation of Hoeffding bound is contrarily relative to n (the quantity of observations) specific node needs to gather enough number of observations so that estimation of € comes out to be less than the value of ΔG.

With R and δ unaltered, the main variable that can be changed in the Hoeffding bound (ε) calculation is the total number of observations (n). At this very moment, ε will decrease with increasing n, as per the assessed information gain getting nearer and nearer to its actual value. A straightforward test permits the decision, with confidence 1-δ that an attribute shows better information gain analyzed than others. This proves to be the main governing principle for Hoeffding tree induction, prompting the accompanying algorithm.

Algorithm for Hoeffding tree induction is given below:

Let HT be a tree with a single leaf (the root)
for all training examples do
Sort example into leaf using HT
Update sufficient statistics in leaf l.
Increment n_1 *mod* n_{min} *the number of examples seen at l.*
if n_1 *mod* n_{min} *= 0 and examples seen at l not all of same class then*
Compute $\bar{G}_l(X_i)$ *for each attribute*
Let X_a*, be attribute with highest* $\bar{G}_l$
Let X_b *be attribute with second-highest* $\bar{G}_l$
Compute Hoeffding bound $\varepsilon = \sqrt{\frac{R^2 ln(1/\delta)}{2n}}$
if $X_a \neq X_b$*, and* $(\bar{G}_l(X_a) - \bar{G}_l(X_b) > \varepsilon$ *or* $\varepsilon < \tau)$ *then*
Replace l with an internal node that splits on X_a
for all branches of the split do
Add a new leaf with initialized sufficient statistics from the split node
end for
end if
end if
end for

Split Confidence: The δ parameter utilized as a part of the Hoeffding bound is one less than the required probability that a right attribute is picked at each point in the decision tree. Since a high probability of accuracy is sought, with likelihood near to one, this parameter is for the most part situated to a little esteem. The estimation of δ is situated to 0.000001

Grace Period: It is computationally excessive to assess the information gain of the attributes after every last preparing case. Given that a solitary case will have little impact on the consequences of the count, it is sensible to sit tight for more samples before re-assessing. The effortlessness period directs what number of samples following the last assessment ought to be found in leaf before returning to the choice. This has the alluring impact of accelerating processing while not significant deviation from exactness. The dominant part of preparing time will be spent redesigning the adequate measurements, a lightweight operation. The most noticeably awful effect is a slow tree development..

Pre-Pruning: It may turn out more important not to split the node by any means. The Hoeffding tree calculation identifies this case by additionally considering the value of no split. A node is just permitted to split when attribute looks adequately superior to anything invalid characteristic, by the same Hoeffding bound test that decides contrasts between different attributes. Pre-pruning does not act as a final decision in the stream setting as compared to the case of batch learning. Nodes are kept from part until it gives the idea that a split will be helpful, so in this sense, the memory management system of nodes can likewise be seen as a type of pre-pruning.

4. Result and Discussion

For experimetation, the complete dataset has been divided into five sub-datasets with 200 records each so as to create the streaming environment because it was not possible to have a real time dataset. So there is a complete dataset with 1000 records and 5 subdatasets with 200 records each. The performance of the proposed approach has been evlauated using accuracy, True positive rate or recall, Precision, F-measure, Kappa statistic, mean absolutete error and root mean squared error. The detalied description of each evlauation metric is given below:

Accuracy: It represents the total number of correctly classified data instances. Higher values of accuracy means a good quality classifier. Accuracy can be formualted defined as:

$$\text{Accuracy}=\frac{\text{TP}+\text{TN}}{\text{TP}+\text{TN}+\text{FP}+\text{FN}} \quad (4)$$

True Positive Rate*:* It is also known as recall or senitivity. It represents the actual number of positively predicted examples which were originally positive and can be defined as:

$$\text{Recall}=\frac{\text{TP}}{\text{TP}+\text{FN}} \quad (5)$$

Precision: Precision represents the actual number of examples which belongs to a class X out of total numbers of examples which are classified in class X and can be formualted as:

$$\text{Precision} = \frac{\text{TP}}{\text{TP}+\text{FP}} \quad (6)$$

F-measure: F-measure is calculated by taking the harmonic mean of Precision and Recall as follows:

$$F - measure = \frac{2*\text{Precision}*\text{Recall}}{\text{Precision}+\text{Recall}} \quad (7)$$

Kappa statistic: It is the measure of agreement or chance corrected agreement between the actual classes and the classification done by the used classifier. The value 0 indicates that the classifier is classifying just by chance and the value greater than 0 indicates that the classifier is actually classifying something.

Mean absolute error (MAE): The mean absolute error represents the average of all the errors occurred while forecasting the classification and at the same time not considering the direction of classification. It is a linear score *i.e*. all the differences in the classification have equal weights which adds up into the error.

Root mean squared error (RMSE): It is just like the MAE but the difference is that MAE is the linear score whereas the RMSE is a degree two or to be precise it is a quadratic score. It is also an average of all the errors in forecasting. Square of all the errors is taken and then the average is taken finally the average is square rooted. As the errors are squared before taking their average that means a higher weight will be assigned to the errors with higher magnitude as compared to the errors with lower magnitude. So, RMSE holds its importance when we want to rule out errors with higher magnitude.

Table 1 shows the experimetal results with and without feature selection approach.

Table 1. Performance Comparison with Static Data

Evaluation Metrics	Without Feature Selection	With Feature Selection
Accuracy	67.00%	**86.56%**
Kappa Statistic	0.0000	**0.6992**
Mean Absolute Error	0.2983	**0.1316**
Root Mean Squared Error	0.3859	**0.2511**
Recall	0.6700	**0.8660**
Precision	0.4490	**0.8660**
F-Measure	0.5390	**0.8660**

Table 2 shows the experimental results of steraming data which are obtained by spliiting the main data into five sub-data with feature selection approach.

Table 2. Performance Comparison with Streaming Data

Evaluation Metrics	Data 1	Data 2	Data 3	Data 4	Data 5	Average
Accuracy	83.92%	86.50%	86.00%	84.50%	89.39%	86.06%
Kappa Statistic	0.6808	0.7212	0.7074	0.6959	0.7798	0.7170
Mean Absolute Error	0.2300	0.1461	0.1972	0.1710	0.1300	0.1749
Root Mean Squared Error	0.3708	0.2809	0.3467	0.2974	0.2560	0.3104
Recall	0.8290	0.8850	0.8975	0.8750	0.9050	0.8783
Precision	0.8900	0.9010	0.9040	0.8950	0.9200	0.9020
F-Measure	0.8584	0.8929	0.9007	0.8849	0.9124	0.8898

So, it can be clearly seen that the classification accuracy of the original dataset is 86.56% (with feature selection) and the mean classification accuracy of five other datasets which are being split for creating the continuous data stream environment is 86.06% and the classification accuracy of the original dataset without feature selection is 67.00%. Next, the performance of the proposed model has been compared with the existing classification approaches such as Neural Network (NN) and k-Nearest Neighbor (k-NN) as shown in table 3. We have mentioned the performance of other classification approaches with feature selection approach. From table 3, it is observed that the VFDT tree with streaming data gives comparable performance with other existing approaches. It shows better performance in terms of kappa statistic, recall, precision and F-measure. The proposed approach takes less computational time as compared to other existing approaches. Therefore, it can be concluded that the proposed stream classification approach is an effective method for link prediction for authorship association in heterogeneous network.

Table 3. Performance Comparison

Evaluation Metrics	Decisiton Tree	VFDT	NN	k-NN
Accuracy	**86.56%**	86.06%	82.04%	78.84%
Kappa Statistic	0.6992	**0.7170**	0.5886	0.4567
Mean Absolute Error	**0.1316**	0.1749	0.2800	0.3120
Root Mean Squared Error	0.2511	0.3104	**0.2485**	0.3365
Recall	0.8660	**0.8783**	0.8580	0.8110
Precision	0.8660	**0.9020**	0.8634	0.8248
F-Measure	0.8660	**0.8898**	0.8607	0.8178
Computaional Time	5.28 sec	**0.46 sec**	6.77 sec	6.35 sec

5. Conclusion

This research work aims at providing the solution of the problem "Link Prediction for Authorship Association in Heterogeneous Network Using Streaming Classification" with the help of an overall architecture which takes DBLP dataset as the input. It combines feature selection algorithm *i.e*. Correlation based Feature Selection (CFS) and classification algorithm *i.e*. Very Fast Decision Tree (VFDT). It can be clearly seen with the help of the detailed results obtained from the experiments done in the dissertation that how the dynamic classification technique has emerged out as the solution of the considered problem. With the advancement in the amount of data/information which is growing at a humongous rate static classification algorithms in future may not be suitable for the classification task, therefore dynamic classification can be seen as a solution for this.

As it is an evident fact that feature selection algorithms reduces the degree of the data to be studied, which means that there may be some features which can play a significant role in the classification task but have been removed because of the feature selection task. To avoid this scenario and to make classification task more useful and accurate, feature weights can play an important role in deciding which feature to keep even after feature selection algorithm has been applied. This will help in identifying all the features which will have any significant impact in the classification task. Though dynamic classification algorithm is used in this research for the classification task but still there is scope of improvement. The dynamic classification algorithm handles data in the form of continuous streams which means it can handle almost infinite amount of data. This algorithm can also be modified to work in parallel or distributed environment. The main advantage of doing this is that it can handle data from many sources apart from handling data from just one source.

References

[1] D. Liben-Nowell and J. Kleinberg, "The link-prediction problem for social networks", Journal of the American society for information science and technology, vol. 58(7), **(2007),** pp. 1019-1031.
[2] M. Al Hasan, V. Chaoji, S. Salem and M. Zaki, "Link prediction using supervised learning". In SDM'06: Workshop on Link Analysis, Counter-terrorism and Security, **(2006)**.
[3] A. Clauset, C. Moore, and M. E. J. Newman, "Hierarchical structure and the prediction of missing links in networks." Nature, vol. 453(7191), **(2008),** pp. 98-101.
[4] C. Wang, V. Satuluri, and S. Parthasarathy, "Local probabilistic models for link prediction." In Seventh IEEE International Conference on Data Mining (ICDM 2007), **(2007)**, pp. 322-331.
[5] R. N. Lichtenwalter, J.T. Lussier, and N.V. Chawla, "New perspectives and methods in link prediction." Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining, **(2010),** pp. 243-252.
[6] V. Leroy, B. B. Cambazoglu, and F. Bonchi, "Cold start link prediction", Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining, **(2010)**, pp. 393-402.
[7] L. Lü and T. Zhou, "Link prediction in complex networks: A survey", Physica A: Statistical Mechanics and its Applications, vol. 390(6), **(2011),** pp.1150-1170.
[8] P. Domingos and G. Hulten, "Mining high-speed data streams", Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining, **(2000),** pp.71-80.
[9] B. R. Prasad and S. Agarwal, "Critical parameter analysis of Vertical Hoeffding Tree for optimized performance using SAMOA", International Journal of Machine Learning and Cybernetics, **(2016),** pp.1-14.
[10] H. Yang, S. Fong, G. Sun, and R. Wong. "A very fast decision tree algorithm for real-time data mining of imperfect data streams in a distributed wireless sensor network." International Journal of Distributed Sensor Networks, **(2012).**
[11] A. M. Hall, "Correlation-based feature selection for machine learning". Diss. The University of Waikato, **(1999).**
[12] E. Budur, S. Lee, and V. S. Kong, "Structural Analysis of Criminal Network and Predicting Hidden Links using Machine Learning", arXiv preprint arXiv:1507.05739, **(2015).**
[13] K. Wei, J. Huang, and S. Fu. "A survey of e-commerce recommender systems." In International Conference on Service Systems and Service Management, **(2007),** pp. 1-5.
[14] T. E. Dao, "Predictive prefetching of web content." U.S. Patent Application 14/311,699, **(2014).**

[15] A. Potgieter, K. A. April, R. J. Cooke, and I. O. Osunmakinde, "Temporality in link prediction: Understanding social complexity". Emergence: Complexity and Organization, 11(1), **(2009),** pp. 69-83.
[16] P. Stoica and K. C. Sharman. "Maximum likelihood methods for direction-of-arrival estimation." IEEE Transactions on Acoustics, Speech and Signal Processing, 38(7), **(1990),** pp. 1132-1143.
[17] M. Sales-Pardo, R. Guimera, A. A. Moreira, and L. A. N. Amaral, "Extracting the hierarchical organization of complex systems." Proceedings of the National Academy of Sciences, vol. 104, no. 39 **(2007),** pp. 15224-15229
[18] H. C. White, S. A. Boorman, and R.L. Breiger, "Social structure from multiple networks. I. Blockmodels of roles and positions." American journal of sociology **(1976),** pp. 730-780.
[19] P. J. Bickel and A. Chen. "A nonparametric view of network models and Newman–Girvan and other modularities." Proceedings of the National Academy of Sciences, vol. 106 (50), **(2009),** pp. 21068-21073.
[20] T. Murata and S. Moriyasu. "Link prediction based on structural properties of online social networks." New Generation Computing , vol. 26, no.3, **(2008),** pp. 245-257.
[21] M. E. J. Newman, "Assortative mixing in networks", Physical review letters, vol. 89, no.20, **(2002)**,pp. 208701.
[22] M. E. J. Newman, "Mixing patterns in networks", Physical review letters, vol. 67, no. 2, **(2003),** pp. 026126.
[23] R. Pastor-Satorras, A. Vázquez, and A. Vespignani, "Dynamical and correlation properties of the Internet", Physical review letters, vol. 87, no. 25, **(2001),** pp.258701.
[24] A. Vázquez, R. Pastor-Satorras, and A. Vespignani, "Large-scale topological and dynamical properties of the Internet", Physical Review E, vol. 65, no. 6, **(2002),** pp. 066130.
[25] M. Girvan and M. E. J. Newman, "Community structure in social and biological networks", Proceedings of the national academy of sciences, vol. 99, no. 12, **(2002),** pp.7821-7826.
[26] R. Guimera, M. Sales-Pardo, and L. AN Amaral, "Classes of complex networks defined by role-to-role connectivity profiles", Nature physics, vol. 3, no. 1, **(2007),** pp.63-69.
[27] J. Reichardt and D. R. White, "Role models for complex networks", The European Physical Journal B, vol. 60, no. 2, **(2007),** pp. 217-224.
[28] L. Getoor, N. Friedman, D. Koller, and B.Taskar, "Learning probabilistic models of link structure", The Journal of Machine Learning Research, vol. 3, **(2003),** pp. 679-707.
[29] A. L. Blum and Pat Langley, "Selection of relevant features and examples in machine learning", Artificial intelligence, vol. 97, no. 1, **(1997),** pp. 245-271.
[30] S. Agarwal and D. Tomar. "A feature selection based model for software defect prediction." International Journal of Advanced Science and Technology , vol. 65, **(2014),** pp. 39-58.
[31] D. Tomar and S. Agarwal, "Hybrid feature selection based weighted least squares twin support vector machine approach for diagnosing breast cancer, hepatitis, and diabetes", Advances in Artificial Neural Systems 2015, **(2015).**
[32] P. Langley, "Selection of relevant features in machine learning", In Proceedings of the AAAI Fall Symposium on Relevance, **(1994),** pp. 1–5.
[33] D. Tomar and S. Agarwal. "Feature selection based least square twin support vector machine for diagnosis of heart disease." International Journal of Bio-Science and Bio-Technology 6, no. 2 **(2014),** pp. 69-82.
[34] W. Siedlecki and J. Sklansky, "On automatic feature selection", International Journal of Pattern Recognition and Artificial Intelligence, vol. 2, no. 2, **(1988),** pp. 197–220.
[35] M. Dash and H. Liu, "Feature Selection for Classification", Intelligent Data Analysis, vol. 1, no. 3, **(1997),** pp. 131– 156.
[36] D. Tomar, D. Ojha, and S. Agarwal. "An emotion detection system based on multi least squares twin support vector machine."Advances in Artificial Intelligence 2014 **(2014).**
[37] D. Tomar and S. Agarwal. "A survey on pre-processing and post-processing techniques in data mining." International Journal of Database Theory and Application, vol. 7, no. 4, **(2014).**
[38] H. H. Hsu, C. W. Hsieh, and M. D. Lu, "Hybrid feature selection by combining filters and wrappers", Expert Systems with Applications, vol. 38, no. 7, **(2011),** pp. 8144-8150.
[39] R. Kohavi, "Wrappers for performance enhancement and oblivious decision graphs". Diss. stanford university, **(1995).**
[40] R. Kohavi and G. H. John, "Wrappers for feature subset selection", Artificial intelligence, vol. 97, no.1 **(1997),** pp. 273-324.

Authors

Harshal Singh. He did his M.tech from Information Technology Division of Indian Institute of Information Technology (IIIT), Allahabad, India under the supervision of Dr. Sonali Agarwal. His primary research interests are Big Data Mining, Data Streaming and Machine Learning.

Divya Tomar. She is a research scholar in Information Technology Division of Indian Institute of Information Technology (IIIT), Allahabad, India under the supervision of Dr. Sonali Agarwal. Her primary research interests are Data Mining, Data Warehousing especially with the application in the area of Medical Healthcare. She has published more than 25 research papers in reputed international Journal and Conferences.

Dr. Sonali Agarwal. Dr. Sonali Agarwal is working as an Assistant Professor in the Information Technology Division of Indian Institute of Information Technology (IIIT), Allahabad, India. Her primary research interests are in the areas of Data Mining, Data Warehousing, E Governance and Software Engineering. Her current focus in the last few years is on the research issues in Data Mining application especially in E Governance and Healthcare. She has published more than 70 research papers in reputed international Journal and Conferences.