Scientometrics
DOI 10.1007/s11192-014-1455-8

Using machine learning techniques for rising star prediction in co-author network

Ali Daud · Muhammad Ahmad · M. S. I. Malik · Dunren Che

Received: 26 May 2014

Abstract Online bibliographic databases are powerful resources for research in data mining and social network analysis especially co-author networks. Predicting future rising stars is to find brilliant scholars/researchers in co-author networks. In this paper, we propose a solution for rising star prediction by applying machine learning techniques. For classification task, discriminative and generative modeling techniques are considered and two algorithms are chosen for each category. The author, co-authorship and venue based information are incorporated, resulting in eleven features with their mathematical formulations. Extensive experiments are performed to analyze the impact of individual feature, category wise and their combination w.r.t classification accuracy. Then, two ranking lists for top 30 scholars are presented from predicted rising stars. In addition, this concept is demonstrated for prediction of rising stars in database domain. Data from DBLP and Arnetminer databases (1996–2000 for wide disciplines) are used for algorithms' experimental analysis.

Keywords Group leader · Classification · Prediction · Rising star · MEMM · CART

A. Daud (✉) · M. S. I. Malik
Department of Computer Science and Software Engineering, International Islamic University, Islamabad, Pakistan
e-mail: ali.daud@iiu.edu.pk

M. S. I. Malik
e-mail: msi_id@yahoo.com

M. Ahmad
Department of Computer Science, Allama Iqbal Open University, Islamabad, Pakistan
e-mail: rana5790@gmail.com

D. Che
Department of Computer Science, Southern Illinois University, Carbondale, IL 62901, USA
e-mail: dche@cs.siu.edu

Published online: 15 October 2014

Introduction

Now-a-days, many online databases such as DBLP or Arnetminer store large numbers of scientific publications. These databanks provide useful information such as the author, venue, publication's title, year and abstract. Additional features such as co-authorship information, co-citation relations in research community may be exploited to facilitate further novel services for online databases.

The academic social networks are typically based on co-authorship and co-citation relationship among researchers and publications. These networks usually have more stable and less dynamic structures as compared to other networks such as social tagging (Tang et al. 2008). There are online services that process the stored information such as Arnet-Miner (Tang et al. 2008) and Microsoft Academic Search.[1] Some portray co-author relationships in a star topology such as Social graph and Instant graph search.[2] However there is a little work done for differentiating different authors and modeling the evolution of author's research profile based over time and progress.

For motivation, the evolutionary behavior of an author is presented in the Fig. 1. In this paper, our goal is to find authors with ascending behavior (rising stars), who have continuous increase in the quantity and quality of their research work in coordination with other scholars (Tsatsaronis et al. 2011). All those persons, who may not be at the top currently or are not experienced, but are capable to be at the top position in their respective fields in future, are referred to as Rising Stars. Rising stars currently have relatively low profiles but may eventually emerge as prominent contributors to the organizations. Searching for rising stars is a new dimension that enables us to find outstanding researchers in Co-author Networks.

There are known research dimensions like finding experts (Daud et al. 2010), research collaborations (Guns et al. 2014), name disambiguation (Huang et al. 2013), citation content analysis (Zhang et al. 2013) and rising stars (Li et al. 2009) in academic social network. Predicting rising stars emerges as a new research area and there is little work done in this regard. An effort (Li et al. 2009) was made to address the problem but limited only to incorporate just author mutual influence and static ranking of publication venues as major features. Later, a methodology (Tsatsaronis et al. 2011) was introduced to address the issue of modeling the dynamics of authors' research profiles. It categorized the authors into different groups using unsupervised learning techniques. The PubRank algorithm was upgraded on StarRank (Daud et al. 2013) by embedding author's contribution based mutual influence and dynamic ranking lists of publications as new features. However there is no work done for predicting rising star using classification techniques.

In this research, we address the problem of finding rising stars by a machine learning approach (classification). Although classification algorithms were already used for expert search (Wang et al. 2013) and different information retrieval tasks (Santos et al. 2013). The goal of this research is to apply classification models (discriminative and generative) to predict future rising stars in the domain of Co-author Network. The outcome of the classifier is "Is a scholar has a potential to become a future rising star or not". Three classes of features (author, venues and co-authorship) are explored and their mathematical formulation is also derived. The main contribution of our work is summarized as follows.

[1] http://academic.research.microsoft.com/.

[2] Co-author Path and Graph in Microsoft Academic Search.

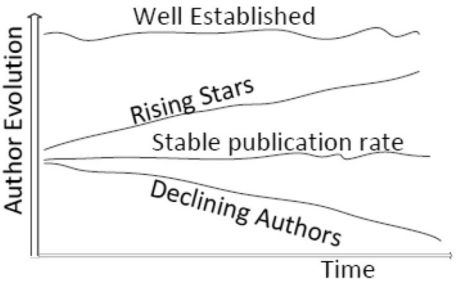


Fig. 1 Four basic author's evolution behavior over time (Tsatsaronis et al. 2011)

1. It is the first attempt that uses supervised machine learning methods for prediction of rising stars. Four famous algorithms are chosen for binary classification of rising stars, although other ML algorithms may also be used. In this work, famous algorithms are selected from wide collection, based on efficient performance and classification accuracy.
2. A set of eleven features are designed on the basis of content and graph information. This feature combination was not considered for prediction of rising stars in previous research.
3. The performance of recommended algorithms is critically analyzed in terms of evaluation metrics and MEMM classifier demonstrates best performance.
4. This novel idea is implemented for rising stars prediction in database domain. It can be implemented for other domains and may be utilized for rising paper prediction.

The remainder of this paper is organized as follows. The related work is presented in second section followed by the problem definition in third section. In fourth section, the applied models (discriminative and generative) are briefly defined and two algorithms are chosen for each type. The detail of evaluation metrics is also presented in fourth section. The description of dataset and proposed feature space with mathematical formulation is presented in fifth section. In sixth section, the methods are implemented and results are analyzed in detail. An application of this framework is also presented in database domain and finally seventh concludes this work.

Related work

As already mentioned, there is little work done for predicting rising stars. So initially we discuss types (discriminative and generative) of machine learning approaches and their applications in social networks.

Application of generative & discriminative classification techniques

"Generative" is a model that formulates joint probability distribution over instances and label sequences. Two algorithms for this model are considered here for classification that is Bayes Network (BN) and Naïve Bayes (NB). BN is already applied for anomaly detection (Mascaro et al. 2014), trust building for electronic markets and communities (Orman 2013), modeling for consensus between expert finding (López-Cruz et al. 2014), simple and complex emotions topic analysis (Ren and Kang 2013) and context adaptive user

interface (Song and Cho 2013). Similar to BN, NB embedded the concept of independence with Bayesian theorem. It employs the concept of conditional probability and successfully implemented for feature subset selection (Bermejo et al. 2014). It is also combined with decision tree for multi-class classification task (Farid et al. 2014). For disambiguation on the affiliations of authors, it was successfully applied (Cuxac et al. 2013). However both (NB and BN) were never applied for rising star classification.

"Discriminative" is based on modeling the dependence of unknown variable on known variable or data. Maximum entropy markov model (MEMM) and CART are used in this research for classification of rising star. Previously, MEMM was successfully implemented for dynamic process monitoring and diagnosis (Li et al.2014), for noise robust speech recognition (Cui et al. 2013) and for Blind separation of non-stationary sources (Gu et al. 2013). CART is a decision tree structure and it is continuously being improved. Examples are multi-labels image annotation (Fakhari and Moghadam 2013) and behavior and credit scoring (Kao et al. 2013). However both (MEMM and CART) were never applied for rising star classification.

Problem definition

Here we define important concepts and provide a formal description of rising star prediction problem.

Rising star Given n training examples $(\mathbf{X}_1, y_1)$, $(\mathbf{X_2}, y_2)\ldots(\mathbf{X_n}, y_n)$, where n is the total numbers of authors, $\mathbf{X_i}$ is a feature vector of author a_i, where $\mathbf{X_i} \in R^m$, m is total number of features and $y_i \in \{-1, +1\}$. To classify whether an author a_i is a rising star or not, the prediction function is defined, which will be learned from training data set.

$$y = F_{RS}(a/\mathbf{X}) \tag{1}$$

where

$$F_{RS}(a/\mathbf{X}) = \begin{bmatrix} \geq 0 & \text{if } y = +1, & \text{rising star} \\ < 0 & \text{if } y = -1, & \text{not rising star} \end{bmatrix} \tag{2}$$

Learning task Our goal is to learn a predictive function $\hat{F}_{RS}$ (.), to predict whether an author a is a rising star or not after a given time period Δt, formally we have

$$\hat{y} = \hat{F}_{RS}(a/\mathbf{X}, \Delta t) \tag{3}$$

We have explored several relevant features to accurately predict a desired label for the corresponding author. In the next section, the mathematical formulation of applied models is presented.

Methods

In this work, two types of classification models are considered to learn the desired predictive function $\hat{F}_{RS}$ (.) and two algorithms are chosen for each model category. In the next section, these algorithms' implementation is examined and results are critically analyzed. The mathematical formulation for these methods is presented next.

Discriminative methods

Maximum entropy Markov model (MEMM)

MEMM or CMM (conditional markov model) combines the functionality of HMM and maximum entropy for labeling of sequential data (McCallum et al. 2000). It basically extends the famous MEC (maximum entropy classifier) with feature that is; unknown parameters are assumed to be connected in a markov chain rather than independent to each others. Given a sequence of observations ($\mathbf{X_1},\ldots,\mathbf{X_n}$), the aim is to tag the observations with the labels ($y_1,\ldots,y_n$) that maximize the conditional probability, defined as

$$P(y_1,\ldots,y_n|\mathbf{X}_1,\ldots,\mathbf{X}_n) = \prod_{t=1}^{n} P(y_t|y_{t-1},\mathbf{X}_t) \tag{4}$$

The transition probabilities are derived from the distribution ($P(y'|\mathbf{X}) = P(y'|y,\mathbf{X})$). Finally for each previous value of label y, the probability of a specific label y′ is calculated as

$$P(y'|\mathbf{X}) = \frac{1}{Z(y,\mathbf{X})}\exp\left(\sum_k \lambda_k f_k\left(y',\mathbf{X}\right)\right) \tag{5}$$

where $Z(y, \mathbf{X})$ is a normalization term that ensures that distribution sum is 1, k is total number of features, λ_k is the factor to be estimated using generalized iterative scaling (GIS) and f_k is real/categorical value feature function.

Classification and regression tree (CART)

CART is basically a non parametric learning approach that results in either regression or classification tree depending variables (features) are either categorical or numeric (Chrysos et al. 2013; Speybroeck 2012). The method of CART contains three steps (Loh 2011).

1. Construction of maximum tree.
2. Selection of right tree size.
3. Classify new data using already constructed tree.

In a simple form, our aim is to predict a response or class y from input vector ($\mathbf{X_1},\ldots,\mathbf{X_m}$). A binary tree is then constructed and a test is performed on each internal node to create a left or right sub branch of tree. This process is repeated until leaf node is constructed. CART solves the following maximization problem at each node.

$$\arg \max_{\mathbf{X}_j \le \mathbf{X}_j^R j=1\ldots m} [I(t_\mathrm{P}) - P_\mathrm{L} I(t_\mathrm{l}) - P_\mathrm{R}(t_\mathrm{r})] \tag{6}$$

where $I(t_\mathrm{p})$ is an impurity function, t_p is parent node; t_l and t_r are left and right child nodes. P_L and P_R are probabilities of left and right child nodes. The gini/towing splitting rules are used for impurity calculation at each node.

Generative methods

Bayes network (BN)

BN is a directed acyclic graph (DAD) with edges as conditional dependencies and nodes as random variables in Bayesian prospective (Mascaro et al. 2014). A probability function is associated with each node of the network for further analysis. Consider a BN comprises n nodes ($\mathbf{X_1}$…. $\mathbf{X_n}$). The joint probability density function of network are calculated as

$$P(\mathbf{X}_1 = x_1, \ldots, \mathbf{X}_{\boldsymbol{n}} = x_n) = P(x_1, \ldots, x_n) = \prod_i P(x_i | x_1, \ldots, x_{i-1}) \tag{7}$$

As we know that in BN, any node's value is conditionally dependent on its parent node. Therefore

$$P(\mathbf{X}_1 = x_1, \ldots, \mathbf{X}_{\boldsymbol{n}} = x_n) = \prod_i P(X_i | \text{Parents}(X_i)) \tag{8}$$

We can write it as

$$P(\mathbf{X}_1 = x_1, \ldots\ldots, \mathbf{X}_{\boldsymbol{n}} = x_n) = \prod_{v=1}^{n} P(\mathbf{X}_v = x_v | \mathbf{X}_{v+1} = x_{v+1}, \ldots, \mathbf{X}_{\mathbf{n}} = x_n) \tag{9}$$

Bayesian network can be used with Bayesian statistics to avoid the problem of data over fitting but it is not bounded to be used with Bayesian statistics (Constantinou et al. 2012).

Naives Bayes (NB)

The NB is a probabilistic classification method that applies naïve hypothesis with Bayes algorithm for every pair of features. It can handle both continuous and categorical independent variables and assumes that features are statistically independent (Ma et al. 2013). Given a class label y and a feature vector $\mathbf{X} = \{x_1, x_2, \ldots, x_m\}$, the Bayes theorem is described as

$$P(y | \mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_3, \ldots, \mathbf{X}_{\boldsymbol{m}}) \propto P(\mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_3, \ldots, \mathbf{X}_{\boldsymbol{m}} | y) P(y) \tag{10}$$

However it assumes that conditional probabilities for independent variables are statistically independent (Chen et al. 2009). Therefore we simplifies the expression into product form as

$$P(y | \mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_3, \ldots, \mathbf{X}_m) \propto P(y) \prod_{k=1}^{m} P(\mathbf{X}_k | y) \tag{11}$$

Finally

$$\hat{y} = \arg\max_y P(y) \prod_{k=1}^{m} P(\mathbf{X}_k | y) \tag{12}$$

Performance evaluation

In this work, the performance of applied classifiers is analyzed by Precision, Recall and F1 evaluation metrics. We mainly used F1 score to examine the effects of different features for rising star classification accuracy and prediction. The mathematical definition for these metrics are described as

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{13}$$

$$\text{Recall} = \text{sensitivity} = \frac{\text{TP}}{(\text{TP} + \text{FN})} \tag{14}$$

$$F1 = 2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}} \tag{15}$$

Data

To evaluate the performance of applied classifiers, the dataset is build by filtering authors' information from DBLP and Arnetminer databases. Initially, about 1.7 million records are crawled for conferences and journals publications. Then distinct authors are identified from a huge collection and relevant properties and characteristics information's are separated. For experimental analysis, we consider authors data for just 5 years (1996–2000) and the goal is to predict future rising stars (future credible researchers) among these authors.

Construction of feature space

In this section, feature set is formed based on contents and graph information. In the next section, four classifiers are trained using this feature information. Then these trained classifiers are evaluated for unseen data. The feature information's are categorized in three types as shown in Table 1. The mathematical formulation and brief description of each feature is also defined here.

Author influence (AI)

If a junior researcher is able to work together with an expert or capable to perform numerous contributions in team work then he/she has bright chances to be a future expert (Li et al. 2009). Consider two authors g and h with 4 and 3 publications respectively. Both are coauthors in 2 publications, the influence of an author to other author can be calculated as

$$\text{Influence}(a_h, a_g) = (a_h, a_g)/\text{Pa}_g = 2/4 = 0.5$$

$$\text{Influence}(a_g, a_h) = (a_g, a_h)/\text{Pa}_h = 2/3 = 0.66$$

where Pa_g and Pa_h are total number of publications for authors a_g and a_h. Hence author a_h influences to a_g with 0.5 score and a_g to a_h with 0.66 score.

Table 1 Features distribution

Author	Venue	Co-author
Author influence	Venue citation	Co-author citations
Author contribution	Venue count	Co-author # of papers
Author contribution based mutual influence	Venue specificity score	Co-author count
Temporal dimension		Co-author venue score

Author contribution (AC)

Author contribution is defined as the amount/value of his participation in a research publication. The contribution of author/co-author is calculated by S index (Sekercioglu 2008) described as.

$$S_k = \frac{1}{k \times H_n} \tag{16}$$

where k is the author rank and n is number of authors. Where $0 \leq S_k \leq 1$ and H_n is calculated as

$$H_n = \sum_{k=1}^{n} \frac{1}{k} \tag{17}$$

Author contribution based mutual influence (ACM)

In this feature, the coauthor relationship is used for the computation of mutual influence. By adding the authors order as they appear in the papers, the corresponding authors are considered as maximum contributors (Daud et al. 2013). The author contribution weight is calculated as

$$\mathrm{ACW}(a_\mathrm{h}, a_\mathrm{g}) = \frac{\left(\sum \mathrm{AC_h} + \sum \mathrm{AC_g}\right)}{\sum \mathrm{PAC_g}} \tag{18}$$

where $\Sigma \mathrm{AC_g}$ is total contribution of author g in all papers and $(\Sigma \mathrm{AC_h} + \Sigma \mathrm{AC_g})$ are authors h and g contribution in coauthored papers. Hence ACW (a_h, a_g) is the author a_h contribution by which he influences author a_g.

Temporal dimension (TD)

The age of an article is analyzed here that is the number of years since its publication and is known as temporal dimension (Yan et al. 2012). If an article is published long before, temporal recency may get positive correlation. The temporal recency is formulated as

$$\mathrm{TRW}(a_\mathrm{g}) = \frac{\sum d(a_\mathrm{g})}{\sum Y(d)} \tag{19}$$

TRW (a_g) is the temporal recency weight of an author a_g, $\sum d(a_\mathrm{g})$ is total number of research articles of author a_g and $\sum Y(d)$ is number of years since its publication.

Venue citations (VC)

In academic social networks, some venues have higher possibility to be more cited than others. Here, the venue impact on citations is investigated and venue citation weight is calculated as

$$\mathrm{VCW}(a_g) = \sum V_x C(d) \tag{20}$$

where VCW (a_g) is the venue citation weight of an author a_g and V_xC (d) for article d is calculated as.

$$V_x C(d) = \frac{\mathrm{VC}a_g}{\sum \mathrm{AVC}} \tag{21}$$

ΣAVC is the sum of all venues' citations and $\mathrm{VC}a_g$ is the specific venue's citations for author a_g.

Venue specificity score (VSS)

Venue specificity or venue similarity is the analysis of venue level and its influence for old and new venues. This venue score is defined as entropy of venue (high rank venues have less entropy and the low rank have high) (Daud et al. 2013). It is formulated as

$$\mathrm{Entropy}(v) = -\sum_{i=1}^{m} w_i \log_2(w_i) \tag{22}$$

where w_i is the $\mathrm{word_i}$ probability in a venue v.

Co-author citations (CC)

If a junior researcher has initially few citations collaborates with senior scholar, then there are more chances for junior scholar to get more citations in collaboration. The co-author citations can be computed as

$$\mathrm{CC}(a_g) = \sum a_b + \sum a_c \sum a_d \tag{23}$$

where CC (a_g) is the co-author citations for author a_g, and ($\Sigma a_b + \Sigma a_c + \Sigma a_d$) is the sum of total papers' citations of the co-authors (a_b, a_c and a_d) of author a_g.

Co-author no of papers (C#), co-author count (CNT) & venue count (VCT)

The total numbers of papers of co-authors are calculated as, first identify all co-authors of a specific paper, sum all papers of each co-author and finally add each coauthor's papers sum to get the result. Co-author count is just to add up the total number of coauthors of a specific author a_g. The venue count calculation is with respect to an author, only count those publication venues where author (a_g)'s papers are published.

Co author venue score (CVS)

Co-author venue score is computed similarly as venue citations, in which we calculate the venue citations of all co-authors for a specific author a_g. The formulation is defined as

$$\mathrm{COVS}(a_g) = \sum \mathrm{VCW}(\mathrm{CO}_{a_g}) \tag{24}$$

where COVS (a_g) is co-author venue score of author a_g and $\sum \mathrm{VCW}(\mathrm{CO}_{a_g})$ shows the sum of venue citations weight of all co-authors for author a_g. Where

$$\mathrm{VCW}(\mathrm{CO}_{a_g}) = \sum V_x C(d) \tag{25}$$

Experimental results

For experiments, publication data from Digital Bibliography and Library Project (DBLP) and Arnetminer databases are used. These databases provide bibliography information for major computer science conferences and journals. This section demonstrates the results of our experiments. The publication data from 1995 to 2000 are used for rising star classification and prediction tasks. First, 44,167 researchers' data are processed and their feature values are computed. From this author's data, total 2,000 instances are selected for experimental evaluation. We build two types of datasets and each set has 1,000 instances. First dataset contains records of top 500 highly cited authors and top 500 with lowest citations. However, second dataset is build with different characteristics. It contains top 500 authors who have highest average relative increase in citations and top 500 authors with lowest average relative increase in citations. The notion for average relative increase in citations (ARIC) is derived similarly as (Tsatsaronis et al. 2011). The ARIC for an author is computed as

$$\mathrm{ARIC} = \max_{i \varepsilon T} \mathrm{Change}_i \times P_{\mathrm{L}} \times \sum_{i \varepsilon T} \mathrm{Change}_i \tag{26}$$

where P_{L} is citations of the last year, T is total no of years and Change_i is the increase in number of citations for current year i. The Change_i is computed as

$$\mathrm{Change}_i = \left(P_i - P_{(i-1)}\right) / P_i \tag{27}$$

where P_i is the number of citations in year i and $P_{(i-1)}$ is the number of citations in last year $(i - 1)$. For label information of rising star (binary classification), the top 500 authors are labeled as rising stars (positive value) and other 500 authors are labeled as not rising stars (negative value) for both data sets. We used fivefold cross validation method to evaluate the performance of classifiers. For each data set, we further divide it into two subsets. First subset contains total 500 instances in which 250 instances are randomly selected from top 500 authors and other 250 instances are selected randomly from other top 500 authors (negative samples). The second subset comprised of total 100 instances with equal number of positive and negative samples. Thus our both data sets and their subsets are chosen by fulfilling the requirement of balance data, resulting in equal number of positive and negative labeled samples. The influence of each feature and its impact on classification accuracy using both data sets is discussed in next subsection and results are attached. The performance of each classifier with all feature combination is presented next. Then rising star score values are calculated for all predicted rising stars and top scholars are ranked in decreasing order of score. Later, these rankings are verified by the current citations and recent status of each scholar in order to realize rising stars predicted potentials.

Feature analysis using classification models

In this section, four suggested methods are implemented for classification of rising stars using both datasets. As described previously, fivefold cross validation technique is chosen for classifiers' training and validation. To analyze the impact of each feature for classification of rising stars, a classifier is trained using each feature that results in total 11 learning cycles and this process is repeated for all classifiers. As each data set has two subsets, the small subset has 100 samples and is partitioned into 10 parts and all classifiers are trained for each partition ranging from 10, 20 …100. The large subset has 500 samples and is partitioned into five parts ranging from 100, 200 … 500. The learning process is performed by WEKA and precision, recall and F1 scores are computed. Here, majority of results are analyzed by F1 score values.

Then mean of F1 score values are computed for each data set. E.g. we computed average of 10 values of F1 score for small subset and vice versa. The average F1 score is plotted against each feature for both data sets as shown in Figs. 2, 3, 4 and 5. The analysis of results for 1st data set that considers total citations as a measure is discussed in this paragraph. The accuracy of classifiers rises when data set size becomes larger. As F1 score by MEMM for Author Contribution feature is 0.74 and 0.84 using small and large subsets respectively. The influence of large data over small data for classification accuracy can be judged by analyzing the F1 score values for other features. The influence of large subset over small subset for classification accuracy can be judged by analyzing the F1 score values for other features as shown in Figs. 2 and 4. However for Venue Count feature, data size has very minor effect on classification accuracy and performance of all classifiers are approximately similar for both subsets.

For 1st data set, the impact of Venue Citations feature for learning rising stars dominates all other features. By using only VC feature, we find 100, 97, 100 and 99 % accuracy with small subset and 100, 99, 100 and 99 % accuracy with large subset by applying MEMM, CART, BN and NB classifiers respectively. Therefore we can infer that this only feature can classify the data more efficiently as compared to other features as shown in Figs. 2 and 4. We find Co-author Venue Score and Venue Count as second and third best features for predicting rising stars for 1st data set. As far classifiers' efficiency, we conclude that MEMM classifier outperforms in predicting rising stars using individual feature.

For 2nd data set that considers average relative increase in citations as a measure, the same experiment is performed and results are shown in Figs. 3 and 5. We did not find any significant increase in classifiers' efficiency when data size increases. Both subsets have almost similar performance for classifiers. Using 2nd dataset, our classifiers' efficiency decrease because this data set has different characteristics and statistics. Again F1 score is considered here for analysis of results and classifiers' performance. The Venue citations feature is found to be the best feature in classify rising stars as it was for 1st data set. But the performance of classifiers decreased as depicted in Figs. 3 and 5. We get 80, 80, 79 and 75 % accuracy by applying MEMM, CART, BN and NB classifiers when only Venue citations feature is incorporated. However Co-author Venue Score and Co-author Citations are found as second and third best features for rising stars prediction task by using 2nd data set. Similarly MEMM classifier again outperforms as it was with 1st dataset and Venue citations feature can help in classification more effectively as compared to other features. Therefore we conclude that the impact of individual feature for classification and performance of classifiers are same for both data sets. This completes the analysis of features for rising stars classification and prediction performance.

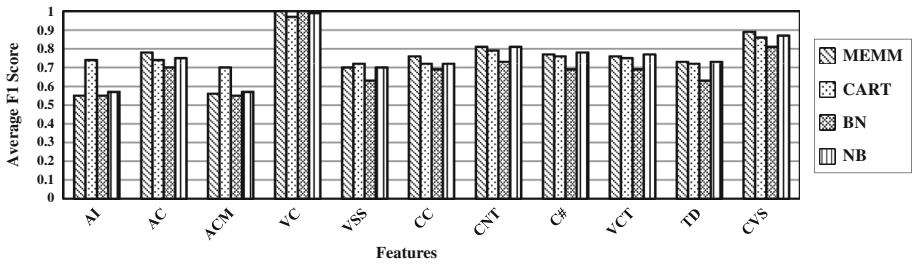


Fig. 2 Impact of individual feature on classification accuracy using 1st data set (100 samples)

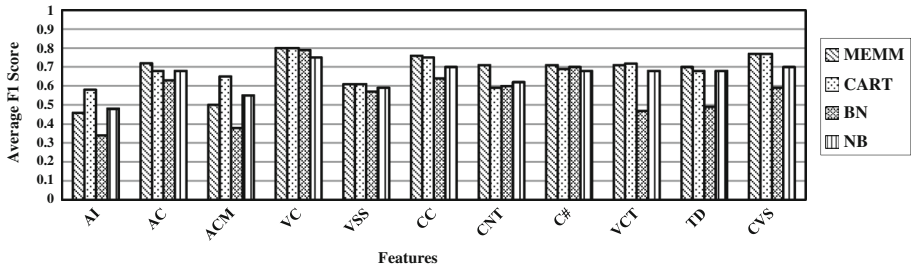


Fig. 3 Impact of individual feature on classification accuracy using 2nd data set (ARIC) (100 samples)

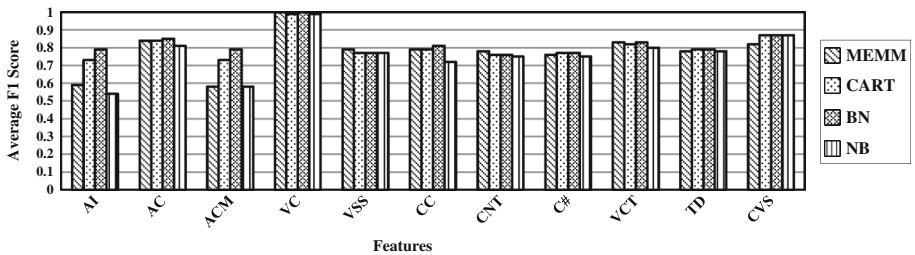


Fig. 4 Impact of individual feature on classification accuracy using 1st dataset (500 samples)

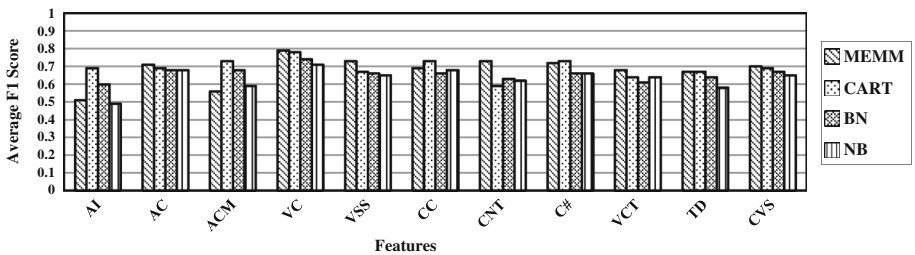


Fig. 5 Impact of individual feature on classification accuracy using 2nd (ARIC) dataset (500 samples)

Classification performance using applied models

In this section, efficiency of four classification methods for learning rising stars classification and prediction task is compared and analyzed in terms of F1 evaluation measures. The same two datasets are used for results generation as it was described in "Feature analysis using classification models" section. Here, classification accuracy of all classifiers improves as data size increases as shown in Figs. 6 and 7. For both datasets, we find the significant improvement in F1 score while upgrading data size for small subsets. For 1st data set, the effect of data size can be judged by analyzing the classification accuracy when data size approaches to 100 samples. At this point, we find max threshold value of F1 score for all classification methods (MEMM, CART, NB and BN) resulting in 100, 99, 92 and 93 % accuracy as shown in Fig. 6. The MEMM classifier demonstrates best performance for rising star classification and prediction task using 1st dataset.

For 2nd dataset, same impact of increase in size of data can be visualized in terms of classifiers' accuracy as shown in Fig. 7 for small subset. The classifiers give best performance and we get maximum F1 score when data size is 100 samples. However for 2nd data set, the accuracy of classifiers reduces as compared to 1st dataset and result in 93, 84, 85 and 86 % for CART, MEMM, NB and BN as shown in Fig. 7. For this dataset, CART classifier is found to be the best method to predict future rising stars.

The performance of four methods is also evaluated using large subset of 1st dataset. The subset size ranges from 100 to 500 samples. It is depicted that by using large subset, each method presents at least 90 % accuracy for rising star prediction as shown in Fig. 8. Therefore accuracy for predicting rising stars is highly correlated to data set size for 1st dataset. We can conclude that efficiency of MEMM again outperforms other methods. However CART algorithm improved its efficiency with large subset and its performance is slightly less than MEMM. Bayesian Network is found to be less effective classifier for finding future rising stars.

The suggested classifiers are also tested using large subset of 2nd data set as shown in Fig. 9. The size of subset ranges from 100 to 500 instances. The performance of classifiers is reduced by using 2nd dataset as compared to performance of classifiers using 1st dataset. The 2nd dataset incorporates average relative increase in citations as a measure for rising stars classification. As data set size increases, the performance of classifiers also increases. Therefore we can infer that accuracy for predicting rising stars is also highly correlated to data set size for 2nd dataset. However for this dataset, CART algorithm outperforms other methods instead of MEMM. But accuracy of MEMM classifier is slightly less than CART. Naïve Bayes is found to be less effective classifier for this dataset as well.

We also evaluated the performance of discriminative and generative methods by using category wise feature combinations. The experiments are performed on both types of datasets and results are shown in Figs. 10 and 11. The first dataset that considers total citations as a measure for rising stars prediction, applied methods give excellent performance when Venue based features are considered for finding future rising stars. Both data subsets are used to verify the effectiveness of applied classifiers by incorporating each category of feature. The average F1 score is computed and we can observe the classifiers' accuracy as depicted in Fig. 10a, b. It results in at least 97 % classification accuracy when Venue based features are employed for predicting future rising stars using both data subsets. We can infer that only Venue based features provide sufficient information to classifiers in order to predict rising stars in co-author network. The prediction accuracy by employing co-author based features is approximately same for both data subsets. MEMM classifier is found to be best method for this data set.

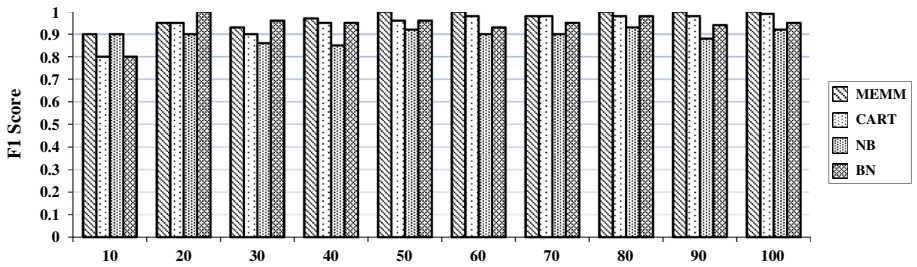


Fig. 6 Classification performance using 1st data set (100 instances)

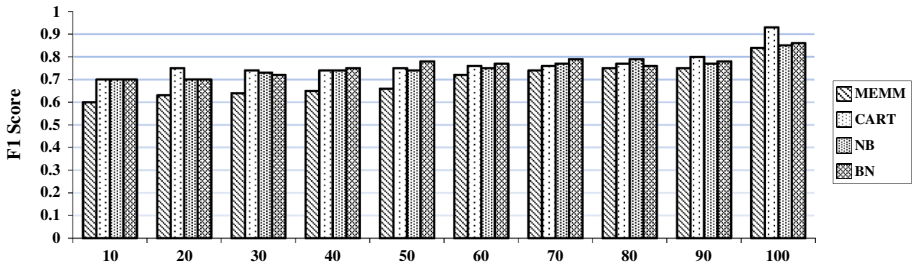


Fig. 7 Classification performance using 2nd data set (ARIC) (100 instances)

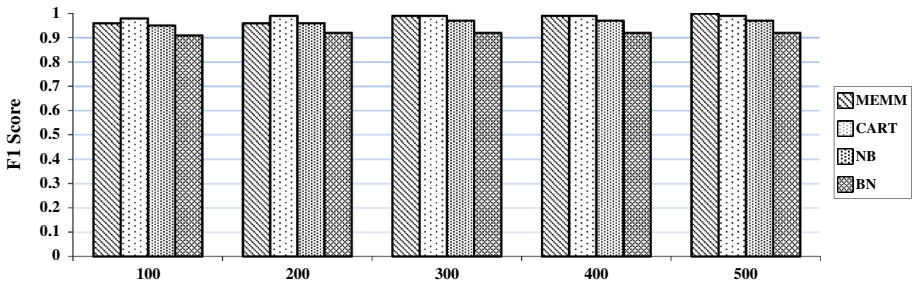


Fig. 8 Classification performance using 1st data set (500 instances)

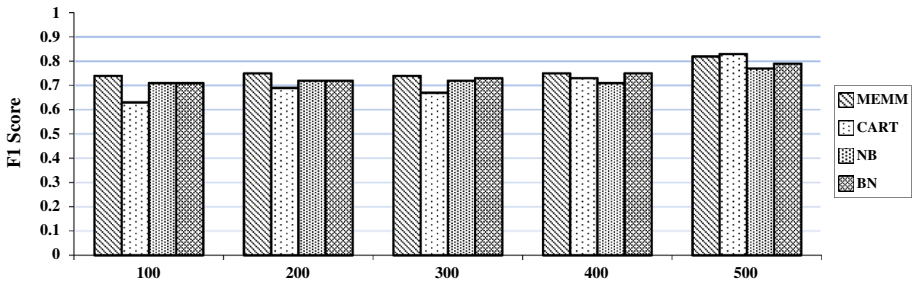
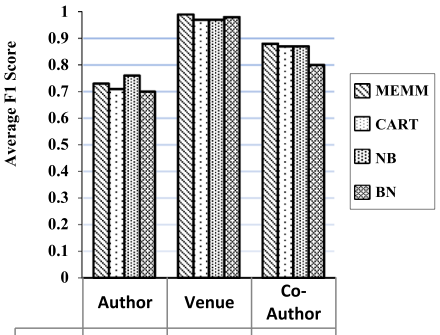
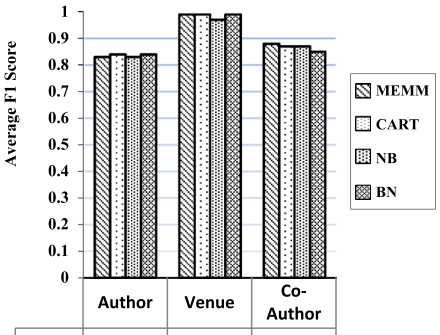


Fig. 9 Classification performance using 2nd data set (ARIC) (500 instances)



	Author	Venue	Co-Author
MEMM	0.73	0.99	0.88
CART	0.71	0.97	0.87
NB	0.76	0.97	0.87
BN	0.7	0.98	0.8

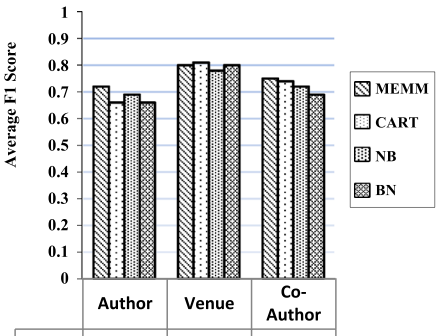
(a) Using small subset (100 samples)



	Author	Venue	Co-Author
MEMM	0.83	0.99	0.88
CART	0.84	0.99	0.87
NB	0.83	0.97	0.87
BN	0.84	0.99	0.85

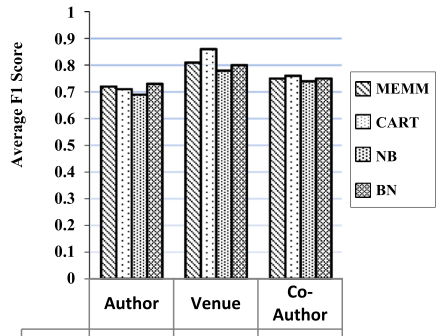
(b) Using large subset (500 samples)

Fig. 10 Category-wise analysis of features using 1st dataset. **a** Using small subset (100 samples), **b** using large subset (500 samples)



	Author	Venue	Co-Author
MEMM	0.72	0.8	0.75
CART	0.66	0.81	0.74
NB	0.69	0.78	0.72
BN	0.66	0.8	0.69

(a) Using small subset (100 samples)



	Author	Venue	Co-Author
MEMM	0.72	0.81	0.75
CART	0.71	0.86	0.76
NB	0.69	0.78	0.74
BN	0.73	0.8	0.75

(b) Using large subset (500 samples)

Fig. 11 Category-wise analysis of features using 2nd (ARIC) dataset. **a** Using small subset (100 samples), **b** Using large subset (500 samples)

For 2nd dataset that considers average relative increase in citations, the same experiment was conducted and results are shown in Fig. 11a, b. Here applied classifiers also give best performance when Venue based features are used for finding future rising stars as in the case of 1st dataset. The three categories of features are analyzed for both data subsets to examine the classification accuracy. The average F1 score shows the impact of each classifier for three categories of features, results in minimum 78 % classification accuracy when Venue based features combination is utilized for both data subsets. We get approximately similar performance for both data subsets when co-author based features are

used. For this data set, CART is found to be best classifier as shown in Fig. 11. MEMM degrades its performance for this data set. However we get better classification accuracy using 1st data set as compared to 2nd data set.

Ranking of predicted rising stars

In this section, ranking of top thirty rising stars are presented for the list of predicted future rising stars for both data sets. The authors are ranked by sorting score in decreasing order as shown in Tables 2 and 3. For 1st data set, the formulation of rising star score is derived in two steps; first values of all features are normalized in the range of 0–1, in second step, all feature values are added except VSS. For all other feature values higher is better so they are added while for VSS lower is better so it is subtracted. The mathematical formulation is given by

$$\text{Rising Star Score} = (AI + AC + ACM + VC + CC + CCT + C\# + VC + CVS + TD) - VSS \tag{28}$$

Then score is sorted in descending order and list of top-30 authors are presented in Table 2. The table presents authors' profiles with full name, career progress, rising star score and total citations up to 2014. The citations of each author prove our hypothesis and we can conclude that all predicted authors have at least 2,000 citations.

However, some authors in the list hold low citations but they are ranked at higher positions as compared to other authors who hold higher citations but ranked at lower positions. E.g. Moses Charikar has 10,215 citations and is ranked at 1st position but Ravi Kumar has 21,258 citations and is ranked at 2nd position as shown in Table 2. The reason is, Moses' co-authors have more publications than Ravi and influence of co-author number of papers, author contribution based mutual influence, author contribution and coauthor citations weights enabled him to be ranked higher than Ravi.

The ranking of top thirty authors for predicted rising stars using 2nd data set is also presented in Table 3. The mathematical formulation of ARIC score is described in Eq. (26). The list contains author's name, current position, ARIC score and total citations up to 2014 similarly as in Table 2. The authors selected by score (average relative increase in citations) presented in Table 3 are different as compared to Table 2 because here highest relative increase in citations is the criteria for ranking of authors. All predicted authors using this data set have at least 6,055 citations. Similarly some authors with low citations are ranked at higher positions as compared to authors with higher citations but ranked at lower positions as in the case of 1st data set. The reason is, those authors who have higher relative increase in citations are ranked at higher positions as compared to those who have low relative increase in citations.

Comparison

Here the comparison of two approaches presented in the experimental section to predict future rising stars is performed. First approach used total citations as a measure and second approach used average relative increase in citations as a measure to predict future rising stars. We have presented a ranking list of top-30 authors developed by each approach. The

Table 2 Ranking of top 30 predicted rising stars using 1st data set

Name	Position/affiliation	Score	Citations
Moses Charikar	Associate Professor, Department of Computer Science Princeton University	2.01	10,215
Ravi Kumar	Senior Staff Research Scientist Google Mountain View, CA	1.77	21,258
Rajeev Alur	Zisman Family Professor Department of Computer and Information Science University of Pennsylvania, USA	1.61	32,007
B. F. Francis Ouellette	Associate Professor, Department of Cell and Systems Biology, University of Toronto	1.51	8,853
Amit Sahai	Professor, Department of Computer Science UCLA, Los Angeles	1.49	15,219
Ee-Peng Lim	Professor School of Information Systems, Singapore Management University	1.45	6,055
Sridhar Rajagopalan	DIMACS Center, IBM Almaden Research Lab K53/802 650 Harry Road	1.45	9,005
Hari Balakrishnan	Professor Department of Computer Science, University of Missouri-Rolla, USA	1.40	64,628
Sudipto Guha	Department of Computer and Information Science University of Pennsylvania	1.36	19,222
Sonia Fahmy	Associate professor, Department of Computer Science Purdue University	1.36	5,940
Mahmut T. Kandemir	Professor, Department of Computer Science and Engineering Pennsylvania State University	1.36	10,007
Barbara A. Rapp	Senior Researcher, National Library of Medicine, 8600 Rockville Pike, Bethesda	1.35	7,834
Byron Dom	Principal Research Scientist, Yahoo! Search Sciences	1.32	9,383
Venkatesan Guruswami	Associate Professor, Computer Science Department Carnegie Mellon University	1.32	7,572
Soumen Chakrabarti	Associate Professor, Computer Science and Engineering, Indian Institute of Technology Bombay	1.32	9,800
Chandra Chekuri	Associate Professor, Algorithms/Theory Group Department of Computer Science University of Illinois Urbana-Champaign	1.31	6,281
David L. Wheeler	National Center for Biotechnology Information, National Institutes of Health, USA	1.28	25,807
Shivkumar Kalyanaraman	Professor, Department of ECSE, Rensselaer Polytechnic Institute (RPI), Dept of CS and Information Technology Center for Pervasive Computing and Networking (CPCN) Networking Laboratory	1.22	3,850
Ian Horrocks	Professor, Computer Science of Oxford University	1.19	23,810
Rajeev Rastogi	Executive Director, Bell Laboratories	1.19	17,765
Thad Starner	Associate Professor, School of Interactive Computing, College of Computing Georgia Institute of Technology	1.14	9,102

Table 2 continued

Name	Position/affiliation	Score	Citations
Wee Keong Ng	Associate Professor, Nanyang Technological University	1.13	2,270
David J. Lipman	Director NCBI, NCBI, NLM, NIH Building 38A, Room 8N807 8600	1.13	46,587
Michael A. Bender	Associate Professor, Department of Computer Science State University of New York at Stony Brook	1.08	3,546
Srinivasan Seshan	Assistant Professor, School of Computer Science Carnegie Mellon University	1.06	12,530
Jeen Broekstra	Researcher Wageningen UR Food & Biobased Research, Technische Universiteit Eindhoven Department of Mathematics and Computer Science Group Information Systems PO Box 513 NL-5600 MB Eindhoven The Netherlands	1.06	3,912
George Karypis	Professor, Department of Computer Science & Engineering, University of Minnesota	1.06	21,565
Gonzalo Navarro	Professor, Department of Computer Science (DCC), Faculty of Physical and Mathematical Sciences, University of Chile	1.05	11,151
Steven D. Gribble	Associate Professor, Department of Computer Science and Engineering University of Washington	1.04	9,341
Erik D. Demaine	Professor, Erik Demaine MIT Computer Science and Artificial Intelligence Laboratory 32 Vassar Street Cambridge, Massachusetts 02139 USA	1.04	10,506

Table 3 Ranking of top 30 predicted rising stars using 2nd data set (ARIC)

Authors	Position/affiliation	ARIC	Citations
Vahid Tarokh	Professor and Senior Fellow of Electrical Engineering, School of Engineering and Applied Sciences	24,521.18	20,763
Ewan Birney	Senior Scientist, European Bioinformatics Institute	17,942.86	15,587
Berthier A. Ribeiro-Neto	Associate Professor, Department of Computer Science, Federal University of Minas Gerais	15,143.10	15,085
Thorsten Joachims	Associate Professor, Department of Computer Science, Cornell University	14,565.40	23,563
Jianbo Shi	Assistant Professor, Department of Computer & Information Science University of Pennsylvania	14,354.38	17,527
Hamid Jafarkhani	Professor Deputy Director Co-Director, Electrical Engineering and Computer Science Center for Pervasive Communications and Computing Networked Systems Program	14,273.97	11,600
Hari Balakrishnan	Professor Department of Computer Science, University of Missouri-Rolla, USA	14,184.12	64,628
Erik L. L. Sonnhammer	Professor, Department of Biochemistry and Biophysics, Stockholm University, Sweden	13,339.94	22,237
B. F. Francis Ouellette	Associate Professor, Department of Cell and Systems Biology, University of Toronto.	13,277.72	6,105
Ian Horrocks	Professor, Computer Science of Oxford University	12,910.78	23,810
Dieter Fox	Associate Professor and Director, Department of Computer Science & Engineering University of Washington	12,801.91	21,321
Nello Cristianini	Professor, Engineering Mathematics and Computer Science University of Bristol	12,171.8357	27,860
Sridhar Rajagopalan	DIMACS Center, IBM Almaden Research Lab K53/802 650 Harry Road	11,713.55	9,005
Steve Lawrence	Senior Staff Research Scientist, Google	10,725.92	12,064
Chris Stauffer	Visiting Scientist, MIT & Founder, Visionary Systems and Research	10,682.67	7,528
Mark Handley	Professor of Networked Systems, Computer Science department at UCL	10,632.87	17,422
Keith A. Crandall	Professor of Biology, Department of Biological Science, George Washington University	10,443.00	14,292
Ravi Kumar	Senior Staff Research Scientist Google Mountain View, CA	9,876.05	21,258
Robert Cooley	Department of Computer Science, University of Minnesota, Minneapolis, MN	9,750.06	6,055
Eckart Zitzler	Assistant Professor for Systems Optimization, Computer Engineering and Networks Laboratory (TIK) ETH Zurich	9,693.30	10,778
Rajeev Rastogi	Executive Director, Bell Laboratories	9,661.19	17,765
David J. Lipman	Director NCBI, NCBI, NLM, NIH Building 38A, Room 8N807 8600	9,497.86	46,587

Table 3 continued

Authors	Position/affiliation	ARIC	Citations
P. Jonathon Phillips	National Institute of Standards and Technology, United States	9,452.69	16,107
George Karypis	Professor, Department of Computer Science & Engineering, University of Minnesota	9,445.45	21,565
Elizabeth M. Belding-Royer	Professor and Vice Chair, Department of Computer Science University of California	9,433.35	21,805
Iftach Nachman	Assistant professor of biology, Tel Aviv University	9,384.23	6,484
Byron Dom	Principal Research Scientist, Yahoo! Search Sciences	9,377.64	9,383
Patrick J. Rauss	Faculty Member, National Institute of Standards and Technology, United States	8,935.28	7,026
Hendrik Blockeel	Professor, Department of Computer Science, Katholieke Universiteit Leuven.	8,463.05	8,003
Sudipto Guha	Department of Computer and Information Science University of Pennsylvania	8,343.67	19,222

Table 4 Ranking of top 30 predicted rising stars for database domain

Name	Position/Affiliation	Score	Citations
Ravi Kumar	Senior Staff Research Scientist Google Mountain View, CA	3.21	21,258
Moses Charikar	Associate Professor, Department of Computer Science Princeton University	3.17	10,215
Barbara A. Rapp	Senior Researcher	2.91	7,834
David L. Wheeler	National Center for Biotechnology Information, National Institutes of Health, USA	2.83	25,807
B. F. Francis Ouellette	Associate Professor, Department of Cell and Systems, Biology, University of Toronto	2.82	8,853
Rohit Goyal	Principal Software Engineer, Axiowave Networks, U.S.A	2.75	1,481
Sonia Fahmy	Associate professor, Department of Computer Science Purdue University	2.68	5,940
Amit Sahai	Professor, Department of Computer Science UCLA, Los Angeles, United States of America	2.67	15,219
Venkatesan Guruswami	Associate Professor, Computer Science Department Carnegie Mellon University	2.66	7,572
Rajeev Alur	Zisman Family Professor Department of Computer and Information Science University of Pennsylvania, USA	2.65	32,007
Sridhar Rajagopalan	DIMACS Center, IBM Almaden Research Lab K53/802 650 Harry Road	2.52	9,005
Evgeni M. Zdobnov	Swiss Institute of Bioinformatics, Department of Genetic Medicine and Development, University of Geneva Medical School	2.43	3,498
Shivkumar Kalyanaraman	Professor, Department of ECSE, Rensselaer Polytechnic Institute (RPI), Dept of CS and Information Technology Center for Pervasive Computing and Networking (CPCN) Networking Laboratory	2.36	3,850
Ian Horrocks	Professor, Computer Science of Oxford University	2.31	23,810
Erik D. Demaine	Professor, Massachusetts Inst. Tech., Lab. for Computer Science	2.30	10,506
Mahmut T. Kandemir	Professor, Department of Computer Science and Engineering Pennsylvania State University	2.30	10,007
Jeen Broekstra	Researcher, Technische Universiteit Eindhoven Department of Mathematics and Computer Science Group Information Systems, Netherlands	2.19	3,912
Michael A. Bender	Associate Professor, Department of Computer Science State University of New York at Stony Brook	2.15	3,546
Soumen Chakrabarti	Associate Professor, Computer Science and Engineering, Indian Institute of Technology Bombay	2.14	9,800
Srinivasan Seshan	Assistant Professor, School of Computer Science Carnegie Mellon University	2.09	12,530

Table 4 continued

Name	Position/Affiliation	Score	Citations
Chandra Chekuri	Associate Professor, Algorithms/Theory Group Department of Computer Science University of Illinois Urbana-Champaign	2.04	6,281
Byron Dom	Principal Research Scientist, Yahoo! Search Sciences	2.03	9,383
Michel C. A. Klein	Assistant professor, Agent System Research group, AI department of the VU University Amsterdam	2.00	4,099
Ayman F. Naguib	Senior Director of Engineering at Qualcomm Research, Silicon Valley	1.98	2,031
Wee Keong Ng	Associate Professor, Nanyang Technological University	1.95	2,270
David J. Lipman	Director NCBI NCBI, NLM, NIH Building 38A, Room 8N807 8600 Rockville Pike Bethesda, MD 20894	1.92	46,587
Bettina Kemme	Associate Professor, School of Computer Science McGill University	1.90	2,804
Fulvio Corno	Associate Professor, Department of Automatica e Informatica (Computer science and automation) of Politecnico di Torino	1.87	2,357
Stephan Tobies	software design engineer, European Microsoft Innovation Center	1.85	3,693
Hari Balakrishnan	Professor, Department of EECS Massachusetts Institute of Technology	1.79	64,628

total citations information of each author in both lists is taken from Arnetminer. The authors presented by Table 2 are sorted by rising star score and those presented in Table 3 are sorted by ARIC score. The rising star score incorporates all features values. So we can deduce that the position of an author in Table 2 is totally based on 11 features values. However the position of an author in Table 3 is based on ARIC value and this ARIC value is calculated for the period 1995–2000 and then authors are ranked by sorting ARIC values in descending order.

Thus we can conclude that there is only one factor (ARIC) used in score calculation for ranking of authors in Table 3. E.g. the author "Vahid Tarokh" has highest average relative increase in citations that's why he is ranked at top position. But author "Moses Charikar" is ranked at top position in Table 2 because he is most influential (AI), best contributor in research (AC ACM), has maximum articles' venue citations (TD, VC), has maximum citations and papers for co-authors (CC, C#, CNT), has maximum venue impact for co-authors (CVS) and has maximum venues where articles are published as compared to other authors in Table 2. However for both cases, we analyze that there are authors who have more citations but are ranked at lower positions than those who are ranked at higher positions but have less total citations. The reason is, some authors would not maintain their current growth in research therefore will get less increase in citations and will show less contribution in collaboration. So both approaches enforce different merits and demerits.

Ranking application of predicted rising stars for database domain

The rising star prediction method can be applied for specific domains. Here it is applied for predicting rising stars in database domain. First, list of database venues are downloaded from Arnetminer. Second, 46,915 publications from these venues for the period 1996–2000 are processed. Third, rising stars ranking of top thirty authors are performed based on rising star score as presented in Eq. (28). Table 4 presents ranking of authors with citations up to 2014, where some predicted rising authors have low citations but ranked at higher positions and vice versa. The features like co-author number of papers, author contribution based mutual influence enable an author to be more famous and highly creditable due to collaborative research activities.

Conclusion

In this paper, discriminative and generative machine learning techniques are used for prediction of rising stars in co-author network. Three classes of features are explored i.e. Author, Co-tuhor and Venue and Venue type is found most effective. MEMM, CART, Naïve Bayes and Bayesian network are chosen for experiment and results analysis. Two types of data sets are made using total citations and average relative increase in citations as measures. It is found that MEMM performs better as compared to other models for total citations and CART performs better as compared to other models for average relative increase in citations. It is also concluded that with the increase of sample size the performance of methods increases especially up to 100 samples though the performance difference from 100 to 500 samples is normal. Ranking lists for top thirty scholars from predicted rising stars are quite trustworthy. At the end, the application of this concept to database domain is also found functional.

For future work, this concept can be applied in other domains e.g. finding rising sellers in online shopping, finding rising cloud service provider in cloud environment, and finding rising auctioneers in online auctions.

References

Bermejo, P., Gamez, J. A., & Puerta, J. M. (2014). Speeding up incremental wrapper feature subset selection with Naive Bayes classifier. *Knowledge-Based Systems, 55*, 140–147.

Chen, J., Huang, H., Tian, S., & Qu, Y. (2009). Feature selection for text classification with Naïve Bayes. *Expert Systems with Applications, 36*(3), 5432–5435.

Chrysos, G., Dagritzikos, P., Papaefstathiou, I., & Dollas, A. (2013). HC-CART: A parallel system implementation of data mining classification and regression tree (CART) algorithm on a multi-FPGA system. *ACM Transactions on Architecture and Code Optimization, 9*(4), 47.

Constantinou, A. C., Fenton, N. E., & Neil, M. (2012). pi-football: A Bayesian network model for forecasting Association Football match outcomes. *Knowledge-Based Systems, 36*, 322–339.

Cui, X., Afify, M., Gao, Y., & Zhou, B. (2013). Stereo hidden Markov modeling for noise robust speech recognition. *Computer Speech & Language, 27*(2), 407–419.

Cuxac, P., Lamirel, J.-C., & Bonvallot, V. (2013). Efficient supervised and semi-supervised approaches for affiliations disambiguation. *Scientometrics, 97*(1), 47–58.

Daud, A., Abbasi, R., & Muhammad, F. (2013). Finding rising stars in social networks. *Database Systems for Advanced Applications (LNCS), 7825*, 13–24.

Daud, A., Li, J., Zhou, L., & Muhammad, F. (2010). Temporal expert finding through generalized time topic modeling. *Knowledge-Based Systems (KBS), 23*(6), 615–625.

Fakhari, A., & Moghadam, A. M. E. (2013). Combination of classification and regression in decision tree for multi-labeling image annotation and retrieval. *Applied Soft Computing, 13*(2), 1292–1302.

Farid, D. M., Zhang, L., Rahman, C. F., Hossain, M. A., & Strachan, R. (2014). Hybrid decision tree and naïve Bayes classifiers for multi-class classification tasks. *Expert Systems with Applications*, *41*(4) Part 2, 1937–1946.

Gu, F., Zhang, H., & Zhu, D. (2013). Blind separation of non-stationary sources using continuous density hidden Markov models. *Digital Signal Processing, 23*(5), 1549–1564.

Guns, R., & Rousseau, R. (2014). Recommending research collaborations using link prediction and random forest classifiers. *Scientometrics,*. doi:10.1007/s11192-013-1228-9.

Huang, S., Yang, B., Yan, S., & Rousseau, R. (2013). Institution name disambiguation for research assessment. *Scientometrics,*. doi:10.1007/s11192-013-1214-2.

Kao, L. J., Chiu, C. C., & Chiu, F. Y. (2013). A Bayesian latent variable model with classification and regression tree approach for behavior and credit scoring. *Knowledge-Based Systems, 36*, 245–252.

Li, Z., Fang, H., & Xia, L. (2014). Increasing mapping based hidden Markov model for dynamic process monitoring and diagnosis. *Expert Systems with Applications, 41*(2), 744–751.

Li, X. K., Foo, C. S., Tew, K. L., & Ng, S. K. (2009).Searching for rising stars in bibliography networks. In *Proceedings of the 14th international conference on database systems for advanced applications* (pp. 288–292).

Loh, W. J. (2011). Classification and regression trees. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 1*(1), 14–23.

López-Cruz, P. L., Larrañaga, P., DeFelipe, J., & Bielza, C. (2014). Bayesian network modeling of the consensus between experts: An application to neuron classification. *International Journal of Approximate Reasoning, 55*(1), 3–22.

Ma, Z., Sun, A., & Cong, G. (2013). On predicting the popularity of newly emerging hashtags in Twitter. *Journal of the American Society for Information Science and Technology, 64*(7), 1399–1410.

Mascaro, S., Nicholso, A. E., & Korb, K. B. (2014). Anomaly detection in vessel tracks using Bayesian networks. *International Journal of Approximate Reasoning, 55*(1), 84–98.

McCallum, A., Freitag, D., & Pereira, F. C. (2000). Maximum entropy Markov models for information extraction and segmentation. In *Proceedings of the seventeenth international conference on machine learning* (pp. 591–598). San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.

Orman, L. V. (2013). Bayesian inference in trust networks. *ACM Transactions on Management Information Systems (TMIS)*, *4*(2), Article No. 7. New York, USA: ACM.

Ren, F., & Kang, X. (2013). Employing hierarchical Bayesian networks in simple and complex emotion topic analysis. *Computer Speech & Language, 27*(4), 943–968.

Santos, R. L. T., Macdonald, C., & Ounis, I. (2013). Learning to rank query suggestions for adhoc and diversity search. *Information Retrieval, 16*(4), 429–451.

Sekercioglu, C. H. (2008). Quantifying co-author contributions. *Science, 322*, 371.

Song, I. J., & Cho, S. B. (2013). Bayesian and behavior networks for context-adaptive user interface in a ubiquitous home environment. *Expert Systems with Applications, 40*(5), 1827–1838.

Speybroeck, N. (2012). Classification and regression trees. *International Journal of Public Health., 57*(1), 243–246.

Tang, J., Zhang, J., Yao, L., Li, J., Zhang, L., & Su, Z. (2008). Arnetminer: Extraction and mining of academic social networks. In *Proceedings of the 14th ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 990–998).

Tsatsaronis, G., Varlamis, I., & Norvag, K. (2011). How to become a group leader? Or modeling author types based on graph mining. *LNCS, 6966*, 15–26.

Wang, G. A., Jiao, J., Abrahams, A. S., Fan, W., & Zhang, Z. (2013). Expert rank: A topic-aware expert finding algorithm for online knowledge communities. *Decision Support Systems, 54*(3), 1442–1451.

Yan, R., Huang, C., Tang, J., Zhang, Y., & Li, X. (2012). To better stand on the shoulder of giants. In *JCDL '12 Proceedings of the 12th ACM/IEEE-CS joint conference on Digital Libraries New York* (pp. 51–60).

Zhang, G., Ding, Y., & Milojevic, S. (2013). Citation content analysis (CCA): A method for syntactic and semantic analysis of citation content. *Journal of the American Society for Information Science and Technology, 64*(7), 1490–1503.