THE 18TH ACM SIGKDD CONFERENCE ON KNOWLEDGE DISCOVERY AND DATA MINING
KDD 2012 BEIJING
Beijing, China August 12-16, 2012

Integrating Community Matching and Outlier Detection for Mining Evolutionary Community Outliers

Manish Gupta UIUC gupta58@illinois.edu

Jing Gao SUNY, Buffalo jing@buffalo.edu

Yizhou Sun UIUC sun22@illinois.edu

Jiawei Han UIUC hanj@illinois.edu

Problem: Find evolutionary community outliers (ECOutliers) (i.e. outliers considering both time and community information).

Examples: (1) An author may change his research area across time while others in his community continue to research in the same area. (2) A stockbroker who suddenly changes his portfolio against his historical investments in some sector even when other similar stockbrokers continue to invest in the same old sector.

State of the art: Community outliers [Gao, KDD10], Temporal outliers [Fox 72], Graph outliers [Aggarwal ICDE11]. We need evolutionary temporal outliers.

Major Idea: We perform cluster matching and outlier detection in a tightly integrated manner. Given two snapshots, we propose a coordinate descent algorithm (1Sμ) to find ECOutliers

Notations: N=#objects; K_1 and K_2:#clusters in X_1 & X_2; P^{NXK_1} and Q^{NXK_2} :Belongingness matrices for X_1 & X_2;$S^{K_1XK_2}$=Correspondence matrix; A^{NXK_2}=Outlierness matrix; μ=Maximum level of outlierness overall

Convex Optimization Formulation

$$\min \sum_{o=1}^{N}\sum_{j=1}^{K_2} \log\left(\frac{1}{a_{oj}}\right)(q_{oj} - \overrightarrow{p_{o.}} . \overrightarrow{s_{.j}})^2$$

subject to the following constraints

$$s_{ij} \geq 0 \quad (\forall i = 1 \dots K_1, \forall j = 1 \dots K_2)$$

$$\sum_{j=1}^{K_2} s_{ij} = 1 \quad (\forall i = 1 \dots K_1)$$

$$1 \geq a_{oj} \geq 0 \quad (\forall o = 1 \dots N, \forall j = 1 \dots K_2)$$

$$\sum_{o=1}^{N}\sum_{j=1}^{K_2} a_{oj} \leq \mu$$

Coodinate Descent Update Rules

$$a_{oj} = \frac{(q_{oj} - \overrightarrow{p_{o.}} . \overrightarrow{s_{.j}})^2 \mu}{\sum_{o'=1}^{N}\sum_{j'=1}^{K_2}(q_{o'j'} - \overrightarrow{p_{o'.}} . \overrightarrow{s_{.j'}})^2}$$

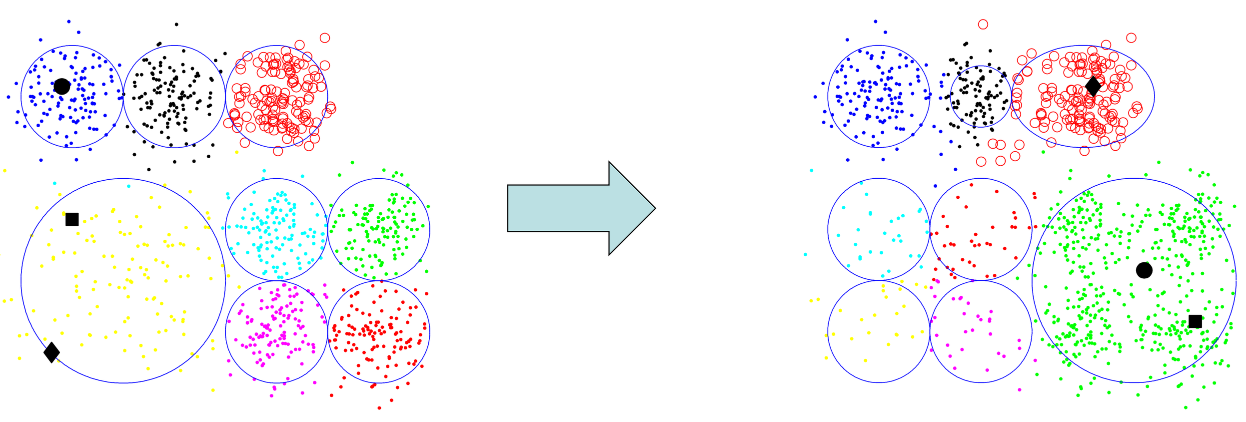
$$s_{ij} = \frac{\sum_{o'=1}^{N} 2\log\left(\frac{1}{a_{o'j}}\right) p_{o'i}[q_{o'j} - \sum_{k=1, k\neq i}^{K_1} p_{o'k} s_{kj}] - \beta_i}{\sum_{o'=1}^{N} 2\log\left(\frac{1}{a_{o'j}}\right) p_{o'i}^2}$$

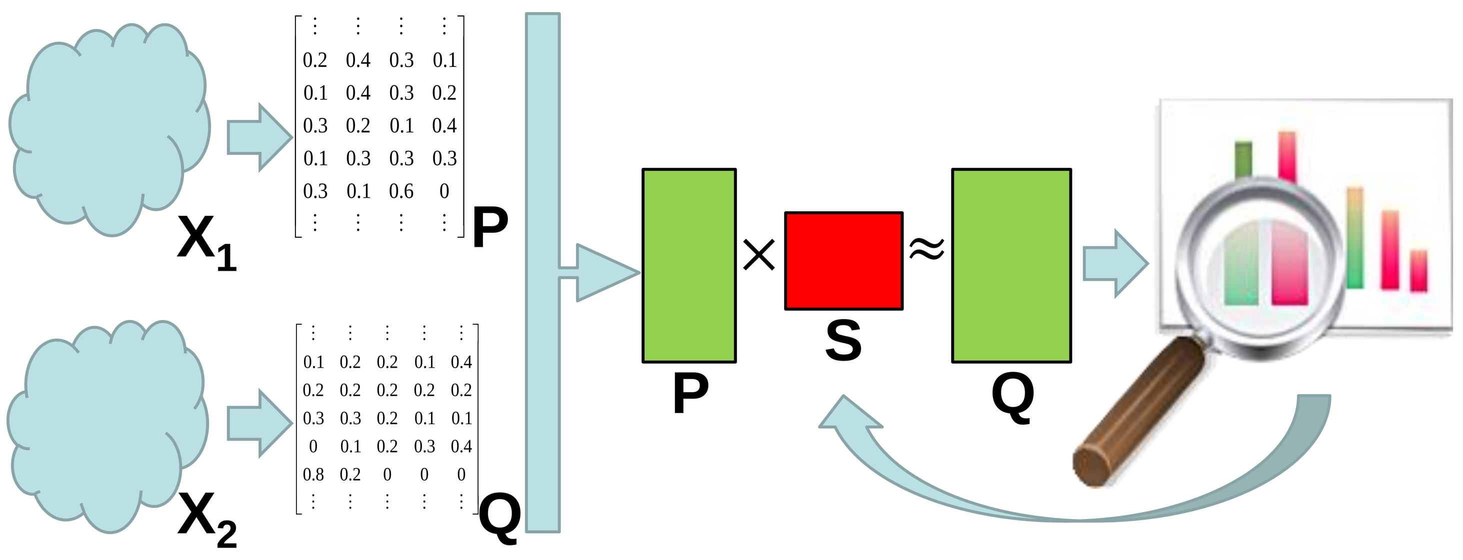
Estimation of Overall Outlierness

$$\mu = \frac{\sum_{o'=1}^{N}\sum_{j'=1}^{K_2}(q_{o'j'} - \overrightarrow{p_{o'.}} . \overrightarrow{s_{.j'}})^2}{\max_{o,j}(q_{oj}^2)}$$

Time Complexity: $O(NK_1^2K_2t)$

Our algorithm is 15% and 33% better than NN and 2S resp.





N	Ψ (%)	SynContractExpand				SynNoEvolution				SynMerge				SynSplit				SynMix			
		NN	2S	1S	1Sμ	NN	2S	1S	1Sμ	NN	2S	1S	1Sμ	NN	2S	1S	1Sμ	NN	2S	1S	1Sμ
1000	1	0.755	0.947	**0.966**	**0.966**	0.832	0.791	0.853	**0.965**	0.72	0.774	0.835	**0.926**	0.786	0.918	0.929	**0.931**	0.606	0.891	0.904	**0.925**
	2	0.729	0.92	0.948	**0.957**	0.812	0.733	0.789	**0.961**	0.702	0.715	0.781	**0.908**	0.779	0.865	0.92	**0.924**	0.675	0.823	0.86	**0.915**
	5	0.71	0.853	0.913	**0.956**	0.726	0.712	0.752	**0.928**	0.645	0.654	0.719	**0.849**	0.697	0.799	0.891	**0.92**	0.631	0.77	0.817	**0.92**
	10	0.619	0.766	0.833	**0.96**	0.657	0.684	0.706	**0.881**	0.58	0.617	0.656	**0.801**	0.63	0.749	0.832	**0.918**	0.594	0.73	0.776	**0.917**
5000	1	0.778	0.945	**0.97**	**0.97**	0.938	0.793	0.848	**0.971**	0.713	0.762	0.801	**0.928**	0.796	0.913	**0.942**	**0.942**	0.691	0.881	0.895	**0.918**
	2	0.756	0.93	0.947	**0.961**	0.864	0.772	0.815	**0.962**	0.677	0.752	0.791	**0.903**	0.768	0.885	0.938	**0.94**	0.646	0.862	0.876	**0.919**
	5	0.689	0.901	0.929	**0.964**	0.742	0.75	0.779	**0.941**	0.626	0.698	0.749	**0.827**	0.689	0.806	0.913	**0.924**	0.608	0.831	0.86	**0.921**
	10	0.622	0.778	0.829	**0.964**	0.656	0.73	0.747	**0.912**	0.579	0.643	0.679	**0.795**	0.624	0.762	0.834	**0.929**	0.593	0.783	0.824	**0.919**
10000	1	0.769	0.949	0.973	**0.974**	0.926	0.807	0.856	**0.974**	0.707	0.788	0.817	**0.933**	0.789	0.938	0.955	**0.96**	0.665	0.882	0.897	**0.921**
	2	0.752	0.937	0.949	**0.963**	0.851	0.788	0.828	**0.964**	0.681	0.762	0.796	**0.898**	0.758	0.898	0.948	**0.951**	0.67	0.869	0.881	**0.916**
	5	0.695	0.9	0.93	**0.964**	0.738	0.763	0.788	**0.951**	0.627	0.719	0.756	**0.826**	0.683	0.807	0.914	**0.922**	0.604	0.847	0.871	**0.919**
	10	0.622	0.771	0.825	**0.965**	0.66	0.753	0.769	**0.926**	0.583	0.645	0.681	**0.795**	0.621	0.769	0.827	**0.934**	0.584	0.812	0.845	**0.917**

Baselines

NN: Comparison with old Nearest neighbors without community matching.

2S: Outlier detection after community matching.

Real Dataset Case Studies

IMDB: Kelly Carlson (I)

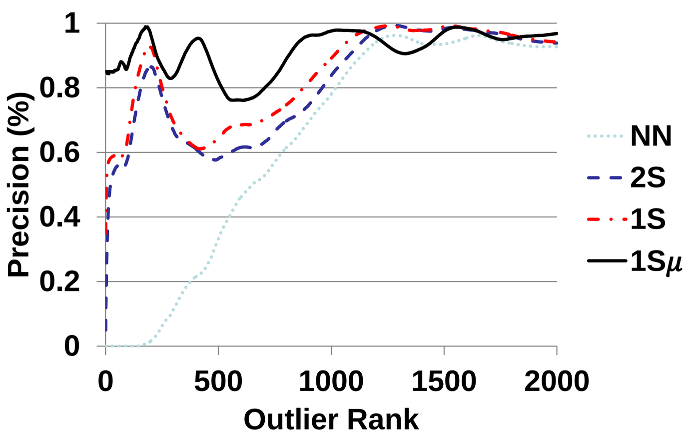
- X_1: Many Sport, Thriller, and Action movies.
- X_2: Many Drama, Music, Reality-TV movies.

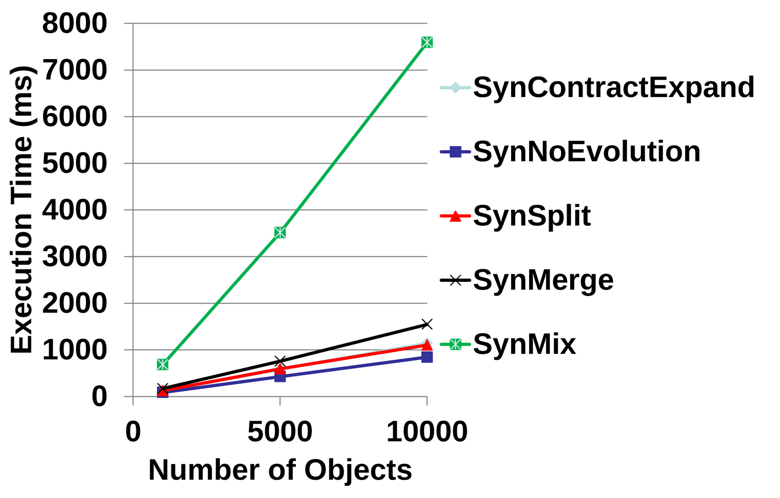
DBLP: Jerome Lang

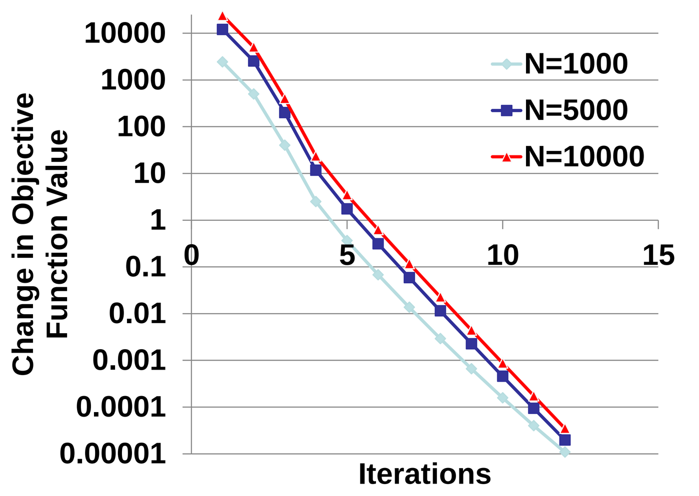
- Stayed back in logic reasoning area (logic, planning, representation, action, uncertainty, propositional)
- His community members moved to data mining and web mining areas (data, retrieval, xml, web, mining).

Features of 1Sμ

- Linear in number of objects
- Converges reasonably fast
- More precise compared to competitive baselines







References

[Aggarwal 11] C. C. Aggarwal, Y. Zhao, and P. S. Yu. Outlier Detection in Graph Streams. In Proc. of the 27th Intl. Conf. on Data Engineering (ICDE), 399–409. 2011.

[Fox 72] A. J. Fox. Outliers in Time Series. Journal of the Royal Statistical Society. Series B (Methodological), 34(3):350–363, 1972.

[Gao 10] J. Gao, F. Liang, W. Fan, C. Wang, Y. Sun, and J. Han. On Community Outliers and their Efficient Detection in Information Networks. In Proc. of the 16th ACM SIGKDD Intl. Conf. on Knowledge Discovery and Data Mining (SIGKDD), 813–822, 2010.

[Long 05] B. Long, Z. M. Zhang, and P. S. Yu. Combining Multiple Clusterings by Soft Correspondence. In Proc. of the 5th IEEE Intl. Conf. on Data Mining (ICDM), 282–289. IEEE Computer Society, 2005.

Summary

We introduced the notion of evolutionary outlier detection with respect to latent evolving communities, **ECOutliers.**

We provided a **coordinate descent algorithm OneStageμ (1Sμ)** that performs community matching across snapshots and ECOutlier detection simultaneously.

Using **three real datasets and multiple synthetic datasets**, we showed the effectiveness of our approach.

Follow-up Work: Community outlier detection for multiple snapshots. Come visit us at ECMLPKDD 12, Bristol, UK.

Paper ID: 110