2015 IEEE International Conference on Data Mining

Post Classification Label Refinement Using Implicit Ordering Constraint Among Data Instances

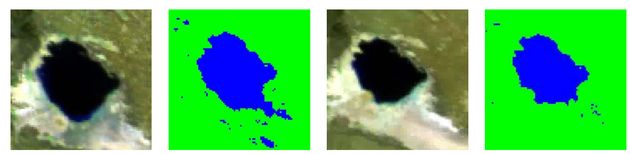
Ankush Khandelwal, Varun Mithal and Vipin Kumar
Department of Computer Science, University of Minnesota
Email: ankush@cs.umn.edu, mithal@cs.umn.edu and kumar@cs.umn.edu

***Abstract*—Classification of instances into different categories in various real world applications suffer from inaccuracies due to lack of representative training data, limitations of classification models, noise and outliers in the input data etc. In this paper we propose a new post classification label refinement method for the scenarios where data instances have an inherent ordering among them that can be leveraged to correct inconsistencies in class labels. We show that by using the ordering constraint, more robust algorithms can be developed than traditional methods. Moreover in most applications where this ordering among instances exists, it is not directly observed. The proposed approach simultaneously estimates the latent ordering among instances and corrects the class labels. We demonstrate the utility of the approach for the application of monitoring the dynamics of lakes and reservoirs. The proposed approach has been evaluated on synthetic datasets with different noise structures and noise levels.**

I. Introduction

In many real world applications, data instances are categorized into binary classes. For example: pass/fail result of a medical test for a certain disease on patients, correct/incorrect answer to a question by students, land/water class type of pixels in the remote sensing image of a region on earth. Moreover, this binary information about the instances can also be collected from multiple sources. For example: multiple pass/fail medical tests for a certain disease on patients, correct/incorrect answer to a set of questions by students, land/water label of pixels from various snapshots of the region taken at different points of time. Due to various data characteristics and classification model limitations these binary class labels are often imperfect. Different post classification refinement methods have been proposed to improve the accuracy of these labels [1], [2]. Post classification refinement methods take these noisy labels as input and aim to improve the accuracy of the labels by making certain assumptions about the application at hand and related data characteristics. In this paper, we present a novel approach for improving label accuracy in scenarios where the instances have a total ordering on them; i.e., if an instance is positive then all instances with ranking higher than that instance will also be positive. If such an ordering is available, it can be used very effectively to correct errors in the labels. However, such an ordering is often not available. In this paper, we present an approach that can effectively correct labels even when the ordering is hidden/unknown. We specifically demonstrate the utility of the approach for the application of monitoring dynamics of lakes and reservoirs from remotely sensed images [3], [4].

We are considering the setting in which individual snapshots of the region have been independently classified with pixels belonging to either land or water class, and hence creates a bipartite grouping of locations. However, bipartite groupings are noisy due to classification inaccuracies. Figure 1 shows a real world example. Figure 1a and 1c show false color composite images of lake Abbe in Brazil at two different timesteps. Figure 1b and 1d show the corresponding classification maps. We can see that both maps have classification errors; hence, there is a need for methods that can exploit some additional information to improve classification performance of these maps.



(a) (b) (c) (d)

Fig. 1: False Color Composites and Classification maps on two timesteps for lake Abbe in Brazil. a) FCC of Jul 1, 2003. b) Classification map of Jul 1, 2003. c) FCC of Aug 10, 2008. d) Classification map of Aug 10, 2008

In this paper, we propose to use the ordering information through elevation for effective correction of class labels. Elevation of a geographical location on earth is its height above a certain fixed point (eg. elevation above sea level). Earth's surface is highly uneven and has various bowl shaped depressions called basins. Water bodies are formed when water fills these basins. Hence, locations inside and around the water have varying elevation.

This elevation information of locations introduces an inherent ordering in the locations. This ordering constraint determines how a water body grows or shrinks. The key idea is; if a location is filled with water, then by laws of physics all the locations in that basin with a lower elevation than the location should also be filled with water. This constraint has been explained in more detail in section III-B. Thus, if we have elevation information then we can detect inconsistent class labels that do not adhere to this physical constraint. Similarly, if we are given perfect class labels, i.e., they perfectly agree to the physics constraint then the growing and shrinking of a lake across various snapshots can be used to extract correct elevation ordering. In other words, by combining information from multiple pure bipartite groupings on instances, instances can be ordered according to their elevation.

Unfortunately, in our setting we are given imperfect class labels (noisy bipartite groupings) and elevation information is also not available. Therefore, we need a way to obtain a good

1550-4786/15 $31.00 © 2015 IEEE
DOI 10.1109/ICDM.2015.149

estimate of elevation (ordering) information from the noisy bipartite groupings coming from different snapshots, so that the estimated ordering can be subsequently used to improve the accuracy of class labels.

In this paper, without loss of generality we assume that higher elevation location appear earlier in the ordering than a lower elevation location. Hence, a shallow location will appear earlier in the ordering through elevation than a deeper location. Under this notation, each pure bipartite grouping can be considered as coarse ordering information where all land pixels appear before all water pixels in the ordering.

The approach proposed in this paper is more general and can be effectively used for any application that has following characteristics -
1. There exists a bipartite grouping of instances and these groupings are available from multiple sources of similar category. For instance, different snapshots of the same region on earth.
2. The bipartite grouping varies between sources due to different characteristics of the sources. For instance, varying extent of a lake across different snapshots.
3. There exists an inherent ordering among instances that can be utilized to detect and subsequently correct the inconsistencies in labels. For instance, the ordering through elevation.

An illustrative example of another application, consider a set of 50 students (instances) taking a chemistry test of 100 questions (sources). Each question has binary (1/0) grading, so each question creates a bipartite grouping of students. The groupings are noisy because some students might have answered a question correctly by chance. The groupings vary across questions because questions are of different difficulty level. If we assume that there exists an ordering among students based on their overall knowledge of chemistry that determines their ability to answer questions, then for any given question it can be estimated which students have answered it correctly by chance based on how students with higher rank have performed on that question. For instance, let's say for a given question, only the top 10 students and the 30^{th} ranking student answered it right. Because of this available ranking information it can be said that the 30^{th} ranking student might have answered this question correctly by chance because the other 19 students, who had better skills than the 30^{th} ranking student, were not able to answer the question. Again, in this scenario as well, bipartite groupings are noisy and the inherent ranking of students is not available.

Given that both pure bipartite groupings and inherent ordering are unknown; in this paper we propose an approach which uses an EM like framework which iterates between estimating inherent ordering from noisy labels and correct labels from the estimated ordering. Specifically, we model the step of estimating total ordering from individual bipartite grouping as a rank aggregation problem. Due to the absence of ground truth, different algorithms have been compared on various synthetic datasets.

II. Related Work

Various post classification label refinement methods have been proposed in remote sensing literature which aim to detect inconsistencies in class label assignments and subsequently correct them by making certain assumptions about the application at hand [5], [1], [6], [7], [2]. These methods tend to perform poorly if the noise in class labels is spatio-temporally auto-correlated, which is the case in our application.

A key element of the proposed approach is to estimate the inherent elevation ordering for correction of inconsistent labels. The problem of obtaining a global ordering of instances by aggregating partial information about the instances has been explored in various areas such as sports, collaborative filtering, meta search and database middle-ware. Various rank aggregation methods have been proposed that aggregate partial information that is available in different forms. There are three major forms of partial information that has been studied in the literature. Partial information can be available in the form of pairwise preference relations [8], rankings on different subsets of instances [9], or as top-k lists from various sources [10]. In our setting, the partial information is available in a new form where bipartite grouping from various sources (snapshots) is available. To the best of our knowledge, this form of partial information has not been studied before in the rank aggregation literature. This bipartite information on instances can also be represented as a set of pairwise preference information between instances. In other words, this form of preference information can be considered as a special case of pairwise preference information from different sources. Hence, methods that use pairwise preference relationships can be potentially used to estimate total ordering [11], [8], [9], [12], [13], [14]. The key difference between the proposed method and these methods is that they do not exploit the extra higher level information that instances have a bipartite grouping in them as well. Bipartite ranking algorithms [15] are also related to this type of information but they work in a supervised setting and also require access to feature space of instances as they learn a ranking/scoring function on the feature space. But in our setting, both feature space and training data is not available.

III. Proposed approach

Now we describe the proposed approach with respect to the application of lake extent monitoring. From here on, the terms snapshot and timestep will be used interchangeably.

A. Input

We are given classification maps of a region across multiple time steps. The matrix $A_{R\times C\times T}$ represents the input data in the 3-dimensional form where $A_{i,j,t} \in \{0,1\}$ represents the class label (0 means land, 1 means water) of a pixel at location (i,j) on the grid at time t. Alternatively, matrix A can be represented as a 2-dimensional matrix $D_{N\times T}$ where N is the total number of locations ($R * C$) and $D_{l,t}$ is the class label of pixel at location index l at time t. From here on, the cells of the grid will be referred to as locations and a (location, timestep) pair defines a pixel. Each row of matrix D represents the temporal sequence of class labels of a location. Since each time step is a binary classification map with noise, each column of D represents a noisy bipartite grouping of the given N instances. Figure 2 shows an illustrative example of these two data representations. For the proposed approach, we will be using the 2-dimensional data representation. Given this input, our aim is to detect and correct inconsistently/incorrectly labelled pixels.

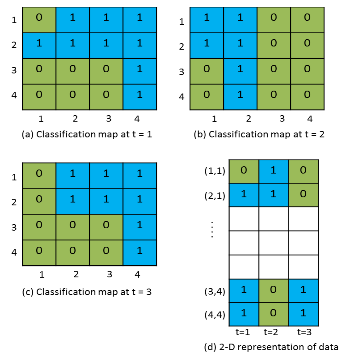


Fig. 2: Illustrative example of 3-D and 2-D representation of the data

B. Physical Constraint through Elevation

In this paper, we propose to use elevation constraint to improve the accuracy of labels. Locations have an inherent ordering based on their elevation. Specifically, if a location l is filled with water then all locations that are deeper (lower elevation) than location l should also be filled with water. Given this constraint through ordering, imperfect labels can be estimated and subsequently corrected. Figure 4a shows a noisy synthetic input matrix D_n with $N = 100$ locations and $T = 200$ timesteps. Blue represents water pixels and green represents land pixels. Each column (timestep) of D_n exhibits an impure pure bipartite grouping. Each row of input matrix D_n represents a temporal sequence of class labels for a location. Figure 4b shows the matrix D_n^π that is obtained by ordering locations in D_n in increasing order of their elevation. Bottom of the matrix represents lowest elevation and top of the matrix represents highest elevation. This ordering information through elevation is referred to as π. We can see that elevation based ordering has reasonably partitioned water and land labels for different timesteps. We can also see the varying nature of bipartite relationships which corresponds to the growing and shrinking of a water body. However, there are some location pairs that are showing physically inconsistent behaviour in some timesteps. For instance, in timestep t a lower elevation (deeper) location k has been labeled as land where a higher elevation (shallow) location i has been labeled as water. Hence, using the elevation ordering these inconsistencies can be detected. However, an issue here is that we still have no idea whether the inconsistency is due to location i being incorrectly labeled as water or location k being incorrectly labeled as land.

C. Estimation of Correct Labels from Elevation

If elevation ordering (π) and size of one partition in bipartite grouping of a timestep t (henceforth referred as θ_t) are available, then inconsistent labels can be estimated and corrected for that timestep. Without the loss of generality, we consider the partition that contains water pixels. For instance, $\theta_t = k$ would mean that for timestep t, bottom k locations in π are filled with water. This would automatically mean that for timestep t, top $N - \theta_t$ locations in π are land where N is the total number of locations.

Unfortunately, in our application θ_t is a latent variable and has to be estimated as well because the data (bipartite groupings) is noisy. Here, the maximum likelihood interpretation of θ_t has been considered. Specifically for a given ordering, that value of θ_t will be selected which leads to least amount of estimated incorrectly labelled locations on that timestep. In order words, that value of θ_t will be chosen which best describes the given bipartite grouping on that timestep. Mathematically, $\hat{\theta}_t$ is estimated as -

$$\hat{\theta}_t = \underset{k \in [0,N]}{\arg\max}\, Acc(k,t,\pi) \tag{1}$$

$$Acc(k,t,\pi) = \sum_{i=0}^{k} \mathbb{1}(D_n^\pi(i,t) == 1) + \sum_{i=k+1}^{N} \mathbb{1}(D_n^\pi(i,t) == 0) \tag{2}$$

The first term in equation 2 measures the agreement of estimated water partition with the noisy water partition, and the second term measures the agreement of the estimated land partition with the noisy land partition. Note that maximum likelihood estimate of $\hat{\theta}_t$ is applicable only when the majority of the labels are correct. In other words, if the majority of the labels are wrong then no method has any hope of correcting the errors. Figure 3 shows an illustrative example demonstrating estimation of $\hat{\theta}_t$. For the given example, $k = 3$ leads to maximum agreement with the noisy input partition with Acc value of 6.

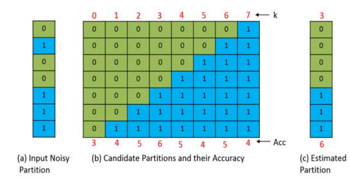
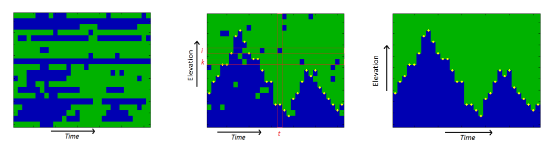


Fig. 3: $\hat{\theta}_t$ estimation example



(a) Noisy Input Matrix D_n **(b)** Ordered Noisy Matrix D_n^π **(c)** Partition Matrix $P_n^{\pi,\hat{\theta}}$

Fig. 4: Synthetic data demonstrating utility of elevation constraint for label correction

Figure 4c shows the matrix $P^{\pi,\hat{\theta}}$ that represents estimated bipartite groupings using equation 1 on D_n^π. Yellow markers on each timestep mark the partition boundary for which estimated partitions showed maximum agreement with noisy partitions. Hence, by definition, for any given timestep locations below yellow markers are labelled as water and locations above the marker are labelled as land. Now, given matrix $P^{\pi,\hat{\theta}}$, we can resolve the ambiguity mentioned in the end of section III-B. Specifically, from $P^{\pi,\hat{\theta}}$ we can say that in D_n^π on timestep t, location i was incorrectly marked as water and not location k.

D. Estimation of Elevation from Perfect labels

In the previous section, we described how accurate elevation information can be used to estimate correct labels. Similarly, accurate class labels (bipartite groupings) can also be used to estimate inherent elevation information. Physically consistent variation in class labels due to dynamics in water body can be used as a proxy to estimate order. Specifically, over any given time period a deeper location will be labeled as water more frequently than a shallow location. This is due to the physics described in section III-B above. Whenever a shallow location is water then the deeper location will strictly be water. But if a deeper location is water then the shallow location may or may not be water depending on the current extent of the water body. Mathematically, elevation of a location is directly proportional to number of time steps in which the location is labeled as water. Specifically, we define the this relationship as -

$$EE_l = T - \sum_{t=1}^{T} \mathbb{1}(D_p(l,t) == 1) \tag{3}$$

where,
EE_l is the estimated elevation of the location l,
T is the number of time steps,

E. Estimation of Elevation from Noisy Labels

Till now, we assumed that either accurate elevation information is available or accurate labels are available. But in our application setting, labels are noisy and even elevation information is not available. EE will only be an approximation of the true hidden elevation information under a noisy labels scenario. In later sections we will show that this a poor approximation of the elevation information.

Therefore, we need a better way to aggregate these noisy bipartite groupings so that a better approximation of inherent ordering (henceforth referred to as $\hat{\pi}$) can be made which will subsequently lead to better label correction capability. Here, we define the objective function that will guide the search for the better approximation of the true ordering. Mathematically, it is defined as -

$$\arg\max_{\hat{\pi}} \sum_{t=0}^{T} Acc(\hat{\theta}_t, t, \hat{\pi}) \tag{4}$$

As mentioned before, Acc is calculated with respect to a given ordering and for a timestep. It measures the agreement between the estimated bipartite grouping and input bipartite grouping. So the above objective function would prefer the ordering that leads to overall better agreement between estimated and input bipartite groupings. In other words, we would prefer an ordering that overall describes the data well. This objective function extends the maximum likelihood interpretation from a single timestep to the whole data. This objective function is very similar to a well known objective function defined as Kemeny optimal aggregation in [16] for finding a global ordering that maximally agrees with input rankings on subset of instances coming from multiple sources. Kemeny optimal aggregations satisfy extended Condercet Criterion which makes them suitable for detecting outliers or inconsistencies in partial rankings [9]. As explained in section III-D, if elevation is given then $\hat{\theta}$ can be estimated but here elevation is also a variable. Hence, in the above function there are two latent variables, $\hat{\theta}$ and $\hat{\pi}$.

Now, to obtain an ordering that best fits the given data or maximizes the above objective function is a NP-Hard problem. Hence, we would need different heuristic solutions to estimate the approximate ordering. Next, we describe the proposed approach (henceforth referred to as *Profile Matching*) that uses an EM like framework that iteratively improves quality of approximate ordering with respect to the above objective function.

F. Proposed Approach: Profile Matching (PM)

As explained in section III-C, if an ordering (π) is available, water levels ($\hat{\theta}_t$) and subsequently correct labels can be estimated for each time step. On the other hand if correct labels are available (section III-D) then elevation information can be accurately estimated. *Profile Matching* iterates between estimating water levels from a given ordering and estimating new ordering from the water levels such that new ordering has improved value of the objective function than the previous ordering. The two steps are explained below -

1) Estimating Water Levels from Ordering: This step is very similar to the step explained in section III-C. The only difference is that here we consider the current estimate of ordering ($\hat{\pi}$) instead of true ordering (π) because it is not available to us. To initialize the process, any random ordering can be provided ($\hat{\pi}^0$). So, water levels at iteration i are calculated as -

$$\hat{\theta}_t^i = \arg\max_{k \in [0,N]} Acc(k, t, \hat{\pi}^{i-1}) \tag{5}$$

2) Estimating ordering from water levels: This is the key step of the approach. First, we describe notion of *location profiles* and *elevation profiles*. As mentioned in section III-B, location profile is basically the temporal sequence of class labels of the given location. Now, consider the matrix shown in 4c. As mentioned earlier $P_n^{\theta,\pi}$ is obtained with respect to a given ordering. This matrix can also be viewed from a different perspective. Each column (timestep) of $P_n^{\theta,\pi}$ represents the estimated bipartite partition for the given column (timestep). On the other hand, each row of $P_n^{\theta,\pi}$ represents the ideal temporal sequence of labels for that row (elevation). To re-emphasize, i^{th} row of $P_n^{\theta,\pi}$ represents the ideal label sequence of i^{th} ranking elevation for the given ordering π. Note that we are distinguishing between location and its elevation. We ideally want a location to be assigned to an elevation such that the location's profile and the corresponding elevation's profile has maximum agreement. For instance, in figure 5, location f has 6 mismatches with the 5th ranking elevation's profile but it has 0 mismatches with 1st ranking elevation's profile. Hence, it makes sense to assign the location f to elevation 1 rather than elevation 5. This transformation of ordering process to the assignment process is the key aspect of the approach. To summarize, we would assign each location to an elevation with which it has maximum agreement. If there are multiple elevations for which a location is showing maximum agreement, then ties

g	0	0	0	1	0	0	1	0
f	1	1	1	1	1	1	1	1
e	1	0	1	1	1	1	1	0
d	0	0	0	0	0	0	1	0
c	0	0	0	0	0	0	0	0
b	1	1	1	1	1	1	1	1
a	0	0	1	1	1	0	1	0

(a) D_n

7	0	0	0	0	0	0	0	0
6	0	0	0	0	0	0	1	0
5	0	0	0	1	0	0	1	0
4	0	0	1	1	1	0	1	0
3	1	0	1	1	1	1	1	0
2	1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1	1

(b) $P_n^{\hat{\pi}^{i-1},\hat{\theta}^i}$

Fig. 5: Assignment operation example

can be broken arbitrarily. Similarly, multiple locations can be associated to a single elevation. Since elevation profiles are already elevation ordered by definition, this assignment process produces partial ordering where instances associated with a higher ranked elevation profile will be ranked higher than instances associated with lower ranked elevation profile. Now to break ties within locations that are assigned to same elevation, we rank them according to their EE rank as there is no other information available for those locations. This gives a new complete ordering of instances.

The algorithm iterates between these two steps until there is no further increase in the value of the objective function.

IV. RESULTS AND DISCUSSION

Next, we describe different algorithms, synthetic datasets and evaluation measures used for comparative analysis. All the codes and synthetic datasets are available at [17].

Traditional spatial majority filter (abbreviated as SS) of window size of 5 pixels and temporal majority filter (TS) of temporal window length of 5 timesteps have been implemented to demonstrate that noise is complex enough that it cannot be removed by trivial majority filters. In order to compare the proposed approach with other ranking based algorithms, two preference based algorithms have been implemented.

Borda Count (BC): Borda Count is a known rank aggregation method mostly used for aggregating votes in an election from a set of voters for a given set of candidates [18]. However, in our application, we do not have total ranking on candidates (locations) from different voters (timesteps) but bipartite information can be converted into score information. Specifically, for a given timestep all water pixels have the same rank and all land pixels have the same rank but a land pixel is ranked higher than all water pixels. Hence, each land pixel gets a score which is equal to the number of water pixels in that timestep. Similarly, each water pixel in that timestep gets a score of 0. In this way total score over all timestep can be calculated and hence can be used for ordering. Also, Borda Count for bipartite groupings turns out to be same as EE explained in section III-D.

Preference Based Ordering (PB): This method uses preference relationship graph to calculate the total ordering among locations [8].

Both of the above ranking based methods returns total order on locations. Once the ranking is obtained, the labels are corrected by using the strategy described in section III-C. To compare the performance of different algorithms, classification error has been used as the evaluation metric.

Next, we describe synthetic data generation process. First, the extent for different timesteps are created such that the dynamics in the lake is physically consistent; i.e., the synthetic water body grows and shrinks according to the predefined inherent ordering of locations. This set of extent maps are the ideal maps that we intend to recover after label correction. Hence, they will be used as ground truth to compare the performance of various algorithms. Five different types of noise structure, namely random noise, spatial noise, temporal noise, spatio-temporal noise and location specific noise are then introduced in the ground truth extents, in varying amounts, to create the noisy datasets that will be provided as input to different algorithms for correction.

Location specific noise is a special case of spatio-temporal auto-correlated noise where the noise is concentrated in few locations. In the application of water monitoring, it happens sometimes that there are some locations that are more noisy that others due to various reasons. So, this noise structure aims to capture that phenomenon.

A. Results

Impact of Noise: The performance of the algorithms depend on the type of noise structure present in the data as well as the amount of noise in the data. Tables I to V show the error rate of different algorithms for different noise structures. The first column of each table reports the amount of noise introduced.

	SS	TS	BC	PB	PM
1 %	1.86	0.53	0.57	0.56	0.05
5 %	2.04	1.73	1.62	1.61	0.24
10 %	2.30	3.63	2.60	2.58	0.60
20 %	3.14	9.89	4.69	4.66	1.86
40 %	17.32	34.19	15.85	15.60	14.54

TABLE I: Random Noise

	SS	TS	BC	PB	PM
1 %	2.51	0.51	0.54	0.54	0.03
5 %	5.38	1.43	1.54	1.54	0.16
10 %	9.32	3.06	2.43	2.41	0.34
20 %	18.22	8.51	4.20	4.14	1.09
40 %	39.21	29.52	16.07	16.11	14.73

TABLE II: Spatial Noise

	SS	TS	BC	PB	PM
1 %	1.85	1.27	0.95	0.94	0.07
5 %	2.02	5.36	4.02	4.01	0.50
10 %	2.25	10.46	6.75	6.74	1.41
20 %	2.97	20.69	11.85	11.79	5.88
40 %	20.25	41.21	35.70	35.56	32.39

TABLE III: Temporal Noise

	SS	TS	BC	PB	PM
1 %	2.53	0.73	0.79	0.78	0.04
5 %	5.43	3.02	2.53	2.56	0.25
10 %	9.25	6.17	4.13	4.10	0.48
20 %	18.01	13.87	6.82	6.77	1.47
40 %	38.82	34.60	20.78	20.92	19.40

TABLE IV: Spatio-temporal Noise

	SS	TS	BC	PB	PM
1 %	2.66	0.89	0.60	0.59	0.08
5 %	5.56	3.37	2.78	2.73	0.38
10 %	9.61	6.75	5.26	5.26	0.89
20 %	18.74	14.91	8.98	8.97	3.41
40 %	39.56	35.57	26.14	26.08	22.97

TABLE V: Location Specific Noise

	SS	TS	BC	PB	PM
RN	3.14	9.90	4.75	4.70	2.04 $\pm$0.11
SN	18.36	8.61	4.04	3.97	1.17 $\pm$0.12
TN	2.99	20.70	11.77	11.70	6.56 $\pm$0.37
STN	18.11	14.27	6.76	6.78	1.24 $\pm$0.11
LN	19.08	15.08	8.74	8.79	4.10 $\pm$0.19

TABLE VI: Impact of initial start on PM

Summarizing above results, performance of TS and SS across different noise structures demonstrate the non trivial nature of noise. Ordering based methods perform significantly better than traditional majority filters. PM shows very good and robust performance across different noise structures and at different noise levels. Total ordering based methods are relatively more sensitive to temporal noise than spatial noise. The performance of BC and PB is very similar because in PB the difference of weights of incoming and outgoing edges in a location [8] are loosely related to the amount of water in the location. Hence both BC and PB have similar characteristics as that of EE defined in section III-D.

Impact of Initial Ordering on PM: Here, we considered datasets with all 3 different types of noise structures with a noise amount of 20%. Table VI shows the mean error rate and standard deviation in error rate when PM was run 100 times; each time with a different random initial ordering for the same given input dataset. The other columns show the error values of other algorithms on the same dataset. As we can see, PM still has better mean error values than other algorithms while having very low standard deviation values. This shows that the algorithm is very robust to the choice of initial ordering.

Importance of Iterations in PM: In this experiment, we analyze the convergence properties of PM. Figure 6 show the percentage improvement in consistency for different datasets with different noise levels. Percentage improvement is calculated with respect to the total improvement achieved by the algorithm from the initial consistency to the consistency in the final iteration. From figure 6, we can say that for all datasets, the first iteration itself contributes the maximum to the overall improvement. However, other remaining iterations do add some improvement in the consistency. Hence, it can be said that there is merit in performing the iterative procedure as it converges quickly and helps in improving consistency.

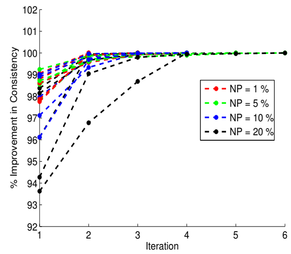


Fig. 6: Percentage Improvement with iterations

B. Limitations

As mentioned earlier, the proposed approach aims to find the parameters that maximize the likelihood of observing the given labels. Due to this, PM expects that the majority of the labels are correct otherwise PM would make poor corrections. Since, the extents should follow the physical constraint, it is necessary that all the water bodies in a region should be impacted in the same way by external forces which cause the water bodies to grow and shrink. As mentioned before, PM and all ordering based methods are more robust to spatial noise than temporal noise.

V. Conclusion and Future Work

In this paper, we proposed a novel approach for improving label accuracy of input class labels in scenarios where there exists an inherent ordering among instances. We demonstrated the utility of the approach for the application of monitoring the extent of lakes and reservoirs. We showed how using elevation constraint can allow development of much more robust algorithms than existing post classification refinement techniques. Similarly, the proposed method demonstrated that it does a better job in exploiting the bipartite information than existing rank aggregation methods. We demonstrated the efficacy of the method of synthetic datasets with varying level of noise complexity.

For future work, the proposed approach can be extended for other scenarios such as working with probabilistic labels rather than binary labels, modelling the quality of sources of bipartite information and uncertainty quantification of correct labels. Since the algorithm converges to a local optimum, it is important to determine bounds on the quality of aggregation with respect to the optimal aggregation. The proposed approach is just one way to effectively utilizing the bipartite information; new methods can be developed for other suitable applications.

VI. Acknowledgments

This research was supported in part by the National Science Foundation Awards 1029711 and 0905581, and the NASA Award NNX12AP37G. Access to computing facilities was provided by the University of Minnesota Supercomputing Institute.

References

[1] D. Liu, M. Kelly, and P. Gong, "A spatial–temporal approach to monitoring forest disease spread using multi-temporal high spatial resolution imagery," *Remote sensing of Env.*, vol. 101, no. 2, pp. 167–180, 2006.

[2] V. Mithal, A. Khandelwal, S. Boriah, K. Steinhauser, and V. Kumar, "Change detection from temporal sequences of class labels: Application to land cover change mapping," in *SIAM International Conference on Data mining, SDM. SIAM*. SIAM, 2013.

[3] F. Sun, W. Sun, J. Chen, and P. Gong, "Comparison and improvement of methods for identifying waterbodies in remotely sensed imagery," *International Journal of Remote Sensing*, vol. 33, no. 21, pp. 6854–6875, 2012.

[4] C. Verpoorter, T. Kutser, D. A. Seekell, and L. J. Tranvik, "A global inventory of lakes based on high-resolution satellite imagery," *Geophysical Research Letters*, vol. 41, no. 18, pp. 6396–6402, 2014, 2014GL060641.

[5] O. Rozenstein and A. Karnieli, "Comparison of methods for land-use classification incorporating remote sensing and gis inputs," *Applied Geography*, vol. 31, no. 2, pp. 533–544, 2011.

[6] X. Chen, J. Chen, Y. Shi, and Y. Yamaguchi, "An automated approach for updating land cover maps based on integrated change detection and classification methods," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 71, pp. 86–95, 2012.

[7] B. Solaiman, R. K. Koffi, M.-C. Mouchot, and A. Hillion, "An information fusion method for multispectral image classification postprocessing," *Geoscience and Remote Sensing, IEEE Transactions on*, vol. 36, no. 2, pp. 395–406, 1998.

[8] W. W. C. R. E. Schapire and Y. Singer, "Learning to order things," in *Advances in Neural Information Processing Systems 10: Proceedings of the 1997 Conference*, vol. 10. MIT Press, 1998, p. 451.

[9] C. Dwork, R. Kumar, M. Naor, and D. Sivakumar, "Rank aggregation methods for the web," in *Proceedings of the 10th international conference on World Wide Web*. ACM, 2001, pp. 613–622.

[10] N. Ailon, "Aggregation of partial rankings, p-ratings and top-m lists," *Algorithmica*, vol. 57, no. 2, pp. 284–300, 2010.

[11] X. Jiang, L.-H. Lim, Y. Yao, and Y. Ye, "Statistical ranking and combinatorial hodge theory," *Mathematical Programming*, vol. 127, no. 1, pp. 203–244, 2011.

[12] E. Zermelo, "Die berechnung der turnier-ergebnisse als ein maximumproblem der wahrscheinlichkeitsrechnung," *Mathematische Zeitschrift*, vol. 29, no. 1, pp. 436–460, 1929.

[13] H. A. David, *The method of paired comparisons*. DTIC Document, 1963, vol. 12.

[14] T. Lu and C. Boutilier, "Learning mallows models with pairwise preferences," in *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, 2011, pp. 145–152.

[15] M.-R. Amini, T.-V. Truong, and C. Goutte, "A boosting algorithm for learning bipartite ranking functions with partially labeled data," in *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2008.

[16] H. P. Young, "Condorcet's theory of voting," *American Political Science Review*, vol. 82, no. 04, pp. 1231–1244, 1988.

[17] "Code and dataset repository," https://github.com/ankushkwal/Profile-Matching-ICDM-2015.

[18] J. A. Aslam and M. Montague, "Models for metasearch," in *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 2001, pp. 276–284.