HELIOSEISMIC AND MAGNETIC IMAGER DATA CLASSIFICATION USING COMBINATORIAL TOPOLOGICAL DYNAMICS

MATEUSZ JUDA

1. Introduction

In this note we present a method for topological features extraction of sampled vector fields. By a sampled vector field we mean a finite set of points in $\mathbb{R}^d$ with vectors attached. Such data arise in a natural way from sampling dynamics. As a real world example we study the data collected by the Helioseismic and Magnetic Imager (HMI) - an instrument designed to study the magnetic field on the surface of Sun [4]. We show that the proposed method significantly outperforms the presently available methods in the HMI solar flare classification task. Our method is general and can be applied to any sampled vector field data, however in this work we present results based only on HMI data.

This note is based on research projects with: Marian Mrozek, Bartosz Zielinski, Tomasz Kapela, Matthias Zeppelzauer.

2. HMI data

The goal of HMI project is to study the relationship between the behavior of the photospheric magnetic field and solar activity. In particular, space weather anomalies are linked to solar flares - a sudden explosion of energy. Solar flares can interfere with satellites and also with equipment such as power utility grids, electronics etc. Predicting solar flares is a challenging task. The recent prediction techniques are based on machine learning (ML) methods. Typically, ML methods for solar flares prediction use 25 numerical characteristics of the magnetic field, the so called data features: total unsigned current helicity, total magnitude of Lorentz force etc.

3. Methodology

We propose to extract features of a sampled vector field using a method based on combinatorial multivector fields [5], a generalization of Forman's combinatorial vector fields [9, 8]. Namely, as a first step we reconstruct dynamics given by a cloud of vectors by building a simplicial complex $\mathcal{K}$ on the point cloud and a combinatorial multivector field $\mathcal{V}$ on $\mathcal{K}$. This way we obtain a graph on the set of all simplices with edges approximating the vector field. We analyze a collection of such graphs using DeepWalk [2] approach which transforms graphs into text documents. Next we use Fasttext [1] to learn embedding of words into $\mathbb{R}^d$, where d is a fixed parameter. Using that embedding we get a representation of the text documents in $\mathbb{R}^d$. The representation gives us a feature vector for each sampled vector field. In the following sections we present more details of the method.

3.1. Multivector fields.

By a *combinatorial dynamical system* on a simplicial complex K (*cds* in short) we mean a multivalued map $F : K \multimap K$, that is a map which sends each simplex in K into a family of simplices in K. The cds F may be viewed as a digraph G_F whose vertices are simplices in K with a directed edge from simplex σ to simplex τ if and only if $\tau \in F(\sigma)$. However, F is more than just the digraph G_F because K, the set of vertices of G_F, is a finite topological space with Alexandrov topology given by the poset of face relation [11].

We construct a cds from a cloud of vectors in two steps. In the first step the cloud of vectors is transformed into a combinatorial multivector field [5]. In the second step, the combinatorial multivector field is transformed into a cds. In order to explain the steps, we introduce some definitions. We say that $A \subset K$ is *convex* if for any $\sigma_1, \sigma_2 \in A$ and $\tau \in K$ such that σ_1 is a face of τ and τ is a face of

Research supported by Polish National Science Center under Maestro Grant 2014/14/A/ST1/00453, and under Sonata Grant2015/19/D/ST6/01215.

σ_2 we have $\tau \in A$. We note that convex subsets of K are precisely the locally closed sets of K (see [6, Sec. 2.7.1, pg 112]) in the Alexandrov topology of K. We define a *multivector* as a convex subset of K and a *combinatorial multivector field* on K (*cmf* in short) as a partition $\mathcal{V}$ of K into multivectors. Given a cmf $\mathcal{V}$, we denote by $[\sigma]_{\mathcal{V}}$ the unique V in $\mathcal{V}$ such that $\sigma \in V$. We associate with $\mathcal{V}$ a cds $F_{\mathcal{V}} : K \multimap K$ given by $F_{\mathcal{V}}(\sigma) := \operatorname{cl}\sigma \cup [\sigma]_{\mathcal{V}}$.

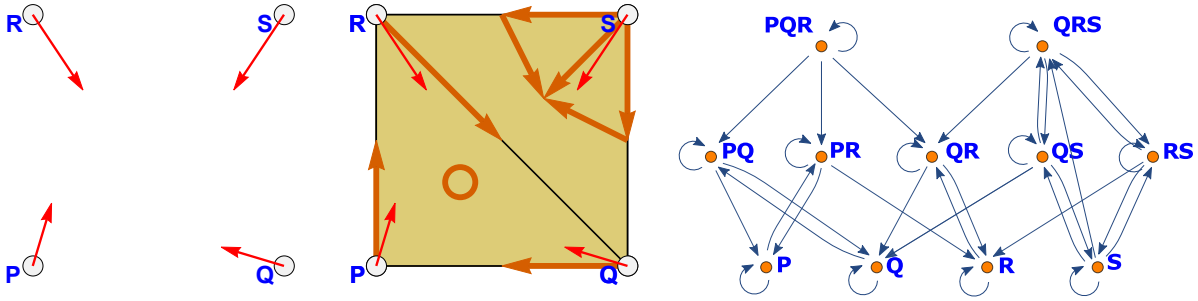

FIGURE 1. Left: A cloud of vectors. Middle: A possible combinatorial multivector field representation of the cloud of vectors. Right: The associated combinatorial dynamical system represented as a digraph.

Figure 1(left) presents a toy example of a cloud of vectors. It consists of four vectors marked red at four points **P**, **Q**, **R**, **S**. One of possible geometric simplicial complexes with vertices at points **P**, **Q**, **R**, **S** is the simplicial complex K consisting of triangles **PQR**, **QRS** and its faces. A possible multivector field $\mathcal{V}$ on K constructed from the cloud of vectors consists of multivectors $\{\mathbf{P}, \mathbf{PR}\}$, $\{\mathbf{R}, \mathbf{QR}\}$, $\{\mathbf{Q}, \mathbf{PQ}\}$, $\{\mathbf{PQR}\}$,$\{\mathbf{S}, \mathbf{RS}, \mathbf{QS}, \mathbf{QRS}\}$. It is indicated in Figure 1(middle) by orange arrows between centers of mass of simplices. Note that in order to keep the figure legible, only arrows in the direction increasing the dimension are marked. The singleton $\{\mathbf{PQR}\}$ is marked with an orange circle. The associated combinatorial dynamical system $F_{\mathcal{V}}$ presented as a digraph is in Figure 1(right). Note that in general K and $\mathcal{V}$ are not uniquely determined by the cloud of vectors.

We denote by $G_{\mathcal{V}}$ the graph obtained from G_F by contracting to a point the vertices in G_F sharing the same multivector.

3.2. **DeepWalk.** In order to analyze a collection of graphs $G_{\mathcal{V}}$ we use DeepWalk [2]. The method is used to analyze graphs as text documents with Natural Language Processing (NLP). Given $G_{\mathcal{V}}$ we generate a set of paths, that is random walks of length not exceeding a fixed k. We assume that for each vertex a word from a vocabulary is given as the vertex label. For a path p we generate a sentence by replacing each vertex on p by its label. A set of such sentences constitutes a text document associated with the set of paths. In this context the order of sentences is not important. For a given set of graphs we consider the documents as a text corpus. Using NLP techniques, in particular Fasttext [1], we learn the representation of words as vectors in $\mathbb{R}^d$ with a fixed d. Each document is represented as the average of its word vectors.

3.3. **Topological vocabulary.** The NLP procedure described above requires a vocabulary in order to assign labels to the vertices. We construct labels which graspe some local, topological properties of the vertex in the vector field. More precisely given a multivector $V \in \mathcal{V}$, that is a vertex in $G_{\mathcal{V}}$, we first define the *label of V at level 0*, denoted $l_0(V)$, as a tuple

$$l_0(V) := (\max_{\sigma \in V} \dim \sigma, |V|, \chi(V)),$$

where $\dim \sigma$ denotes the dimension of simplex σ, $|V|$ stands for the cardinality of V, and $\chi(V)$ is the Euler characteristic of V. We define *label of V at level d* , denoted $l_d(V)$, as a tuple

$$l_d(V) := (l_0(V), \operatorname{sorted}(\{l_0(u) \mid u \in N_d^+(V)\}), \operatorname{sorted}(\{l_0(u) \mid u \in N_d^-(V)\})),$$

where $N_d^+(V)$ (resp. $N_d^-(V)$) are sets of vertices in the forward (resp. backward) distance from V not bigger than d.

As an example we consider the multivector field and the graph G_F presented in Figure 1. Figure 2 presents the associated graph on multivectors $G_{\mathcal{V}}$. Table 1 presents step by step calculations of the labels at level 1.

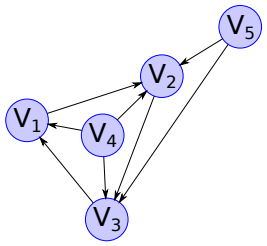

FIGURE 2. $G_{\mathcal{V}}$ graph for the example presented in Figure 1.

V	simplices of V	$l_0(V)$	$N_1^+(V)$	$N_1^-(V)$	$l_1(V)$
V_1	{**P**, **PR**}	$(1,2,0)$	$\{V_2\}$	$\{V_3, V_4\}$	$((1,2,0),[(1,2,0)],[(1,2,0),(2,1,1)])$
V_2	{**R**, **QR**}	$(1,2,0)$	$\{V_3\}$	$\{V_1, V_4, V_5\}$	$((1,2,0),[(1,2,0)],[(1,2,0),(2,1,1),(2,4,0)])$
V_3	{**Q**, **QP**}	$(1,2,0)$	$\{V_1\}$	$\{V_2, V_4, V_5\}$	$((1,2,0),[(1,2,0)],[(1,2,0),(2,1,1),(2,4,0)])$
V_4	{**PQR**}	$(2,1,1)$	$\{V_1, V_2, V_3\}$	$\emptyset$	$((2,1,1),[(1,2,0),(1,2,0),(1,2,0)],\emptyset)$
V_5	{**S**, **RS**, **QS**, **QRS**}	$(2,4,0)$	$\{V_2, V_3\}$	$\emptyset$	$((2,4,0),[(1,2,0),(1,2,0)],\emptyset)$

TABLE 1. Step by step calculation of labels at level 1 for the example presented in Figure 1 and Figure 2

4. RESULTS

To evaluate our method we use a data set proposed in [3]. The data set provides 823 HMI magnetograms. The state-of-the-art methods extract from each magnetogram 13 real number characteristics. Additionally, for each magnetogram we know a flare class (B, C, M, and X) according to the maximum magnitude of flares generated in the approaching 24 hours. Our goal is to find an ML model for the flare class prediction based on the megnetograms.

We use randomly selected 70% of the data as a training set, and the rest as a test set. We transform the magnetograms from the training set into text documents and create a model of the artificial language described above. Then, for each magnetogram (training and test), we create a new feature vector using the word embeddings. We present results obtained with the following parameters:

- level of labels is $k = 4$;
- for each label l we select randomly 50% of vertices v in $G_{\mathcal{V}}$, such that $l_k(v) = l$;
- for each selected vertex we generate a random walk which begins at v and a random walk which ends at v, both of length 20;
- the dimension of the word embeddings is 40.

To compare the state-of-the-art feature vector with the new one we compare classification metrics for LinearSVC [7] and AdaBoostClassifier [10] from sklearn python library. We provide classiffiers scores in Table 2. We observe that the features based on the proposed word embeddings always are significantly better than the state-of-the-art features. We emphesize that the proposed method outperforms state-of-the-art for the test set.

	Classifier	test	training
proposed feature vector	LinearSVC	0.898	0.881
	AdaBoostClassifier	0.846	0.994
state-of-the-art feature vector	LinearSVC	0.417	0.392
	AdaBoostClassifier	0.663	0.918

TABLE 2. Classifiers scores on test and training data sets.

References

[1] A. Joulin, E. Grave, and P. Bojanowski, T. Mikolov. Bag of Tricks for Efficient Text Classification, *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers* (2017), 427–431.

[2] B. Perozzi, R. Al-Rfou, and S. Skiena. Deepwalk: Online learning of social representations, *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining* (2014), 701–710.

[3] C. Liu, N. Deng, J. Wang, H. Wang. Predicting solar flares using SDO/HMI vector magnetic data products and the random forest algorithm. *The Astrophysical Journal*, 843(2), (2017), 104.

[4] M. Bobra, X. Sun, J. Hoeksema,M. Turmon, Y. Liu, K. Hayashi, G. Barnes, K. Leka. The Helioseismic and Magnetic Imager (HMI) Vector Magnetic Field Pipeline: SHARPs – Space-Weather HMI Active Region Patches. *Solar Physics*, 289(9), 3549–3578.

[5] M. Mrozek. Conley-Morse-Forman theory for combinatorial multivector fields on Lefschetz complexes, *Foundations of Computational Mathematics*, **17**(2017), 1585–1633. DOI: 10.1007/s10208-016-9330-z.

[6] R. Engelking. General Topology, *Heldermann Verlag*, Berlin, 1989.

[7] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin. LIBLINEAR: A library for large linear classification. *Journal of Machine Learning Research* 9 (2008), 1871–1874.

[8] R. Forman. Combinatorial vector fields and dynamical systems, *Mathematische Zeitschrift* **228** (1998), 629–681.

[9] R. Forman. Morse theory for cell complexes, *Advances in Mathematics*, **134** (1998), 90–145.

[10] T. Hastie, S. Rosset, J. Zhu, H. Zou. Multi-class adaboost. *Statistics and its Interface*, (2009), 2(3), 349-360.

[11] T.K. Dey, M. Juda, T. Kapela, J. Kubica, M. Lipinski, M. Mrozek. Persistent Homology of Morse Decompositions in Combinatorial Dynamics. arXiv preprint arXiv:1801.06590 (2018).

(Mateusz Juda) Division of Computational Mathematics, Institute of Computer Science and Computational Mathematics, Faculty of Mathematics and Computer Science, Jagiellonian University

E-mail address: mateusz.juda@uj.edu.pl