Verbal Characterization of Probabilistic Clusters using Minimal Discriminative Propositions

Yoshitaka Kameya, Satoru Nakamura, Tatsuya Iwasaki, Taisuke Sato
Graduate School of Information Science and Engineering, Tokyo Institute of Technology
Ookayama 2-12-1, Meguro-ku, Tokyo 152-8552, Japan
Email: {kameya,nakamura,iwasaki,sato}@mi.cs.titech.ac.jp

***Abstract*—In a knowledge discovery process, interpretation and evaluation of the mined results are indispensable in practice. In the case of data clustering, however, it is often difficult to see in what aspect each cluster has been formed. This paper proposes a method for automatic and objective characterization or "verbalization" of the clusters obtained by mixture models, in which we collect conjunctions of propositions (attribute-value pairs) that help us interpret or evaluate the clusters. The proposed method provides us with a new, in-depth and consistent tool for cluster interpretation/evaluation, and works for various types of datasets including continuous attributes and missing values. Experimental results exhibit the utility of the proposed method, and the importance of the feedbacks from the interpretation/evaluation step.**

***Keywords*-knowledge discovery, interpretation, evaluation, clustering, mixture models, emerging patterns**

I. Introduction

In a knowledge discovery process, interpretation and evaluation of the mined results are indispensable in practice. In the case of data clustering, however, it is often difficult to see in what aspect each cluster has been formed, only from a list of the instances in the cluster. Visualization is a natural way for understanding things, and particularly in text clustering, Hotho et al. applied formal concept analysis with Hasse diagrams to visualize the similarity and dissimilarity among the obtained clusters [1]. On the other hand, since there would generally be a physical limitation or a high implementational cost in visualization, we would rather like to "verbalize" the clusters, i.e. we associate an intuitive descriptive label (or a set of such labels) with each cluster. Additionally it seems desirable that the labels are chosen objectively and automatically from the clusters. So far, there have been only a few labeling methods, e.g. LabelSOM [2] and Mei et al.'s automatic labeling for topic models [3]. CLIQUE [4] also has a similar motivation to ours in that it performs hyper-rectangular clustering and at the same time produces comprehensible descriptions of the obtained clusters.

In this paper, we propose a new labeling method that associates conjunctions of propositions (attribute-value pairs), called *propositional labels*, with the clusters obtained by mixture models. To find these propositional labels objectively and automatically, we conduct an Apriori-style breadth-first search for minimal propositional labels that discriminate the cluster of interest from the others. Due to these features, as we will see later, the proposed method can provide us with a new, in-depth and consistent tool for cluster interpretation/evaluation. It is also notable that, unlike the previous attempts, the proposed method is fully applicable to various types of datasets including continuous attributes and missing values. Another novel contribution of this paper is to show empirically the importance of the feedbacks from the interpretation/evaluation step in achieving a reasonable clustering result.

The rest of this paper is structured as follows. In Section II, we describe the details of the proposed method. Section III then reports the experimental results. Finally, we conclude the paper in Section IV. A full description of the proposed method, experimental results and related work is described in [5].

II. Proposed method

A. Preliminaries

First, we roughly introduce some terminology and notation. Suppose that we have a dataset $\mathcal{D}$ of N instances which are described by m discrete attributes $A_1, A_2, \ldots, A_m$, and refer to each instance by $\boldsymbol{a} = (a_1, a_2, \ldots, a_m)$, where a_j is a value of A_j. Also we write $\mathcal{V}(A_j)$ as the set of possible values of A_j. We now introduce a propositional label (or a label, for short) "$X_1 = x_1$" $\wedge$ "$X_2 = x_2$" $\wedge \cdots \wedge$ "$X_n = x_n$" such that $\{X_1, X_2, \ldots, X_n\} \subseteq \{A_1, A_2, \ldots, A_m\}$, X_i and $X_{i'}$ are distinct ($i \neq i'$), and $x_i \in \mathcal{V}(X_i)$. In a probabilistic context, $p($"$X_1 = x_1$" $\wedge \cdots \wedge$ "$X_n = x_n$"$) = p(X_1 = x_1, \ldots, X_n = x_n)$ holds. Also, $p(Z = z, \ldots)$ for a random discrete variable Z and its value z is generally abbreviated as $p(z, \ldots)$ if the context is clear. Furthermore, a label "$X_1 = x_1$" $\wedge \cdots \wedge$ "$X_n = x_n$" is simplified as $\boldsymbol{x} = (x_1 \wedge \cdots \wedge x_n)$ or $\boldsymbol{x} = (x_1, \ldots, x_n)$. For a label $\boldsymbol{x}$ and its subconjunction (resp. proper subconjunction) $\boldsymbol{x}'$, we denote $\boldsymbol{x}' \subseteq \boldsymbol{x}$ (resp. $\boldsymbol{x}' \subset \boldsymbol{x}$).

B. Overview

In this paper, we consider probabilistic clustering based on a simple mixture model called a naive Bayes model. A naive Bayes model has a latent class variable C taking on the identifiers $\{1, 2, \ldots, K\}$ of K clusters, and represents a simple joint distribution: $p(C = k, A_1 = a_1, \ldots, A_m = a_m) = p(C = k) \prod_{j=1}^{m} P(A_j = a_j \mid C = k)$, or equivalently $p(k, \boldsymbol{a}) = p(k) \prod_j p(a_j \mid k)$. Here the probabilities $p(k)$ and $p(a_j \mid k)$ are treated as the model parameters. Given a dataset $\mathcal{D}$ of instances and the number K of clusters, we do:

1) Estimate the parameters in a model $p(k, \boldsymbol{a})$ from $\mathcal{D}$.

2) Assign the most probable class $k^*(\boldsymbol{a}) = \mathrm{argmax}_{1 \leq k \leq K} p(k \mid \boldsymbol{a})$ to each instance $\boldsymbol{a}$ based on the estimated parameters. The k-th cluster $\mathcal{C}_k$ is then formed as a set of instances $\boldsymbol{a}$ such that $k^*(\boldsymbol{a}) = k$.
3) Find propositional labels $\boldsymbol{x}$ that characterize well each cluster $\mathcal{C}_k$.

In the first two steps, we perform clustering, and the third step is called *labeling*. As is well-known, the first step is realized by the EM (expectation-maximization) algorithm. From the second step, clustering can be casted as an unsupervised classification task, and we call $p(k \mid \boldsymbol{a})$ the *(class) membership probability* of an instance $\boldsymbol{a}$. The last step will be described in the next two sections.

C. Characteristic propositional labels

1) Relevance scores: To choose suitable propositional labels $\boldsymbol{x}$ of a cluster $\mathcal{C}_k$ objectively and automatically, we introduce a scoring function that measures how relevant $\boldsymbol{x}$ and $\mathcal{C}_k$ are. Previously, several *relevance scores* have been proposed in various statistical/data-mining tasks (e.g. [6], [7] for comprehensive surveys). In this paper, we choose $p(k \mid \boldsymbol{x})$ as the relevance score for two reasons on intuitiveness for the end users. First, we can of course interpret $p(k \mid \boldsymbol{x})$ as discriminative probabilities, by which we classify an instance satisfying $\boldsymbol{x}$. As mentioned in Section II-B, clustering is performed based on the membership probabilities $p(k \mid \boldsymbol{a})$, which are a special case of $p(k \mid \boldsymbol{x})$. The second reason is more practical: $p(k \mid \boldsymbol{x})$ is inherently normalized (i.e. $0 \leq p(k \mid \boldsymbol{x}) \leq 1$). From this nature, we can use a threshold $r \in (0, 1]$ and is commonly applied to all clusters, to filter out $\boldsymbol{x}$ such that $p(k \mid \boldsymbol{x}) < r$.

2) Minimality: For two labels $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ such that $p(k \mid \boldsymbol{x}_1) \geq r$ and $p(k \mid \boldsymbol{x}_2) \geq r$ for some threshold r, we favor $\boldsymbol{x}_1$ over $\boldsymbol{x}_2$ if $\boldsymbol{x}_1 \subset \boldsymbol{x}_2$, because the longer one may have some redundant information which hinders us from understanding the cluster. In other words, we would like to have only *minimal* labels. Minimality is also taken into account in the literature on emerging pattern mining (e.g. [8]).

3) Model-based computation of relevance scores: We have introduced several relevance scores which are based on probabilities. In most of the previous work, these probabilities are directly estimated from a given dataset $\mathcal{D}$ of instances. For example, membership probabilities are estimated as $\hat{p}(k \mid \boldsymbol{x}) = |\{\boldsymbol{a} \in \mathcal{C}_k \mid \boldsymbol{x} \subseteq \boldsymbol{a}\}| \ / \ |\{\boldsymbol{a} \in \mathcal{D} \mid \boldsymbol{x} \subseteq \boldsymbol{a}\}|$. In our method, on the other hand, relevance scores are computed from the model parameters via the joint distribution (Section II-B). This model-based approach has a couple of advantages. First, as seen later, we can efficiently compute the scores, exploiting the conditional independence in the model, without scanning the whole dataset $\mathcal{D}$. In many cases, the space for the model parameters is much smaller than the dataset. The second advantage is that the model parameters are well-abstracted data as long as the model fits to $\mathcal{D}$, and there would be less chance to be affected by noise. Finally, there is a positive side-effect that we need not care about missing values in $\mathcal{D}$ since we only use the parameters estimated by the EM algorithm.

4) Selecting characteristic propositional labels: Now based on the discussions above, we define *characteristic propositional labels*, which characterize well the obtained clusters. A propositional label $\boldsymbol{x}$ of the cluster $\mathcal{C}_k$ is characteristic iff:

1) $p(k \mid \boldsymbol{x}) \geq r$,
2) $p(\boldsymbol{x}) \geq s_{\mathrm{global}}$,
3) $p(\boldsymbol{x} \mid k) \geq s_{\mathrm{local}}$, and
4) There is no $\boldsymbol{x}' \subset \boldsymbol{x}$ that satisfies 1∼3 above,

where r, s_{global} and s_{local} are user-specified thresholds.

While frequent pattern mining algorithms run based on the guide from the threshold for $p(\boldsymbol{x} \mid k)$ or $p(\boldsymbol{x})$, we treat the first and the fourth conditions as the primary filters. The second and the third conditions are introduced to remedy the problem that we often obtain unintuitive characteristic labels with very low global/local support, and also to reduce the burden in the exhaustive search for characteristic labels, which will be described in the next section. So currently we do not consider to put a tight restriction on global/local support.[1]

D. Exhaustive search for characteristic propositional labels

All possible propositional labels form a lattice, and on this structure, we conduct an Apriori-style breadth-first search for the entire set of characteristic labels for each cluster.[2] We take a breadth-first style because, as seen later, it is easier to check the minimality of characteristic labels in a breadth-first style, and because we do not necessarily need long characteristic labels that are difficult to read. It should also be remarked that our search algorithm can deal with both discrete and continuous attributes in a uniform fashion.

III. Experiments

In this section, we show an experimental result with the flags dataset.[3] Since the flags dataset does not contain the class information, we explore a plausible number of clusters by characteristic labels together with a Bayesian score for model selection called the Cheeseman-Stutz score [9]. We tried 1,000 re-initializations in the EM algorithm not to get trapped into unwanted local optima. The flags dataset contains the details of 194 national flags, originally described by 30 attributes. In this experiment, we focused on the clusters of national flags grouped on their visual aspects, and hence non-visual attributes (landmass, zone, area, population, language and religion) were removed in advance. Since the class information is not given in this dataset, we first estimated the number $\hat{K}$ of clusters by the Cheeseman-Stutz score, as a starting point. Another point in this dataset is that discrete attributes and continuous attributes are mixed. That is, all of eight integer attributes (e.g. the number of circles in the flag) were treated as continuous attributes. We used a threshold $r = 0.75$ for $p(k \mid \boldsymbol{x})$ and conducted a greedy pruning, whose description is omitted.

[1] For example, $s_{\mathrm{local}} = 1/(|\mathcal{D}|/K) = K/|\mathcal{D}|$, which implies that each of equally-sized clusters should contain at least one instance. We often set a small value (e.g. $1/|\mathcal{D}|$) to the threshold s_{global}, so that s_{global} is negligible.

[2] The details of the search algorithm are presented in the full paper [5].

[3] The flags dataset is provided at the UCI ML Repository (http://archive.ics.uci.edu/ml/). For three other experiments, refer to the full paper [5]. Clustering is done by NBCTK available at http://sato-www.cs.titech.ac.jp/nbctk/ .

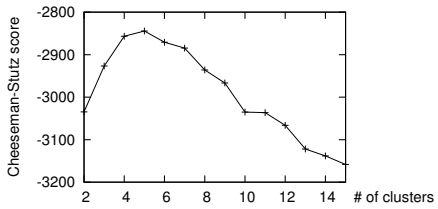

Fig. 1. The Cheeseman-Stutz scores with various numbers of clusters.

TABLE I
THE CHARACTERISTIC LABELS FOR $\mathcal{C}_1, \ldots, \mathcal{C}_6$ IN THE FLAGS DATASET.

labels for $\mathcal{C}_1$	$p(k\|\boldsymbol{x})$	$p(\boldsymbol{x}\|k)$
#saltires=1	1.000	0.900
topleft=white $\wedge$ #quarters=1	0.817	0.622
stripes=0,1,2 $\wedge$ #quarters=1	0.827	0.540
:	:	:
labels for $\mathcal{C}_2$	$p(k\|\boldsymbol{x})$	$p(\boldsymbol{x}\|k)$
#bars=1,2,3,4	0.782	0.800
labels for $\mathcal{C}_3$	$p(k\|\boldsymbol{x})$	$p(\boldsymbol{x}\|k)$
#circles=1,2 $\wedge$ #crosses=0	0.781	0.540
#circles=1,2 $\wedge$ #quarters=0	0.781	0.540
black=T $\wedge$ #circles=1	0.766	0.225
:	:	:
labels for $\mathcal{C}_4$	$p(k\|\boldsymbol{x})$	$p(\boldsymbol{x}\|k)$
#crosses=1 $\wedge$ #saltires=0	0.810	0.81003
#crosses=1 $\wedge$ #quarters=0	0.829	0.81002
#crosses=1 $\wedge$ #sunstars=0	0.751	0.720
#circles=0 $\wedge$ #crosses=1	0.768	0.640
:	:	:
labels for $\mathcal{C}_5$	$p(k\|\boldsymbol{x})$	$p(\boldsymbol{x}\|k)$
#bars=0	0.803	0.900
#circles=0	0.752	0.900
#crosses=0	0.755	0.600
#quarters=0	0.752	0.400
triangle=T	0.889	0.240
:	:	:
labels for $\mathcal{C}_6$	$p(k\|\boldsymbol{x})$	$p(\boldsymbol{x}\|k)$
#saltires=0 $\wedge$ #quarters=1	0.960	0.360
topleft=blue $\wedge$ #quarters=1	0.875	0.320

Fig. 1 shows the curve of the Cheeseman-Stutz score with various numbers of clusters, and we have $\hat{K} = 5$ as a peak of this curve. We further continued to compute characteristic labels with the number K of clusters being around $\hat{K}$, and found that readable characteristic labels are obtained when $K = 6$. Table I presents these labels.[4] The shortest characteristic label for the cluster $\mathcal{C}_1$ says that the national flags in $\mathcal{C}_1$ (and none in the other clusters) have one saltire (diagonal cross). A typical example of such flags is the Union Jack, and actually many flags in $\mathcal{C}_1$ have one quartered section (i.e. #quarters=1) for the Union Jack. Similarly, the clusters $\mathcal{C}_2$ and $\mathcal{C}_3$ contain the flags with vertical bars and with circles, respectively. The label (#saltires=0 $\wedge$ #quarters=1) for $\mathcal{C}_6$ distinguishes $\mathcal{C}_1$ and $\mathcal{C}_6$, and similarly the labels (#crosses=1 $\wedge$ #saltires=0) and (#crosses=1 $\wedge$ #quarters=0) for $\mathcal{C}_4$ jointly work for distinguishing $\mathcal{C}_4$ from $\mathcal{C}_1$ and $\mathcal{C}_6$, where #crosses indicates the number of upright crosses. Indeed, $\mathcal{C}_6$ contains the flag of the United States, and $\mathcal{C}_4$ contains the flags of several Scandinavian countries (note that the Union Jack also contains upright crosses). From the labels for $\mathcal{C}_5$, one may see that $\mathcal{C}_5$ is a cluster of miscellaneous flags. On the other hand, when the number K of clusters is set at $\hat{K} = 5$, the clusters $\mathcal{C}_2$ and $\mathcal{C}_3$ are merged into one cluster, whose characteristic labels are not so intuitive as in Table I. These results imply that a plausible number of clusters can be determined by interactively consulting characteristic labels, with a help from model selection techniques, and clearly show how the feedbacks from the interpretation/evaluation step contribute in knowledge discovery.

[4]Since each continuous attribute A_j is originally an integer attribute, a proposition "$\alpha < A_j \leq \beta$" (assume here that α and β are not integers, for simplicity) was translated back into "$A_j = \lceil\alpha\rceil, \lceil\alpha\rceil + 1, \ldots, \lfloor\beta\rfloor$" in Table I. Non-minimal labels produced by this translation were then removed.

IV. CONCLUSION

In this paper, we proposed a new labeling method that associates propositional labels (conjunctions of attribute-value pairs) with the clusters obtained by mixture models, to help us interpret or evaluate the clusters. As shown in the experimental results, the proposed method finds a set of intuitive descriptive labels that characterize well or "verbalize" the clusters. The proposed method is fully applicable to various datasets including continuous attributes and missing values, and can be a new, in-depth and consistent tool for cluster interpretation/evaluation. Experimental results show that the feedbacks from the interpretation/evaluation step can play an important role for achieving a reasonable clustering result.

ACKNOWLEDGMENTS.

The authors would like to thank Toshihiro Kamishima and anonymous reviewers for helpful comments on related work. This work is supported in part by Grant-in-Aid for Scientific Research (No. 20240016) from MEXT of Japan.

REFERENCES

[1] A. Hotho, S. Staab, and G. Stumme, "Explaining text clustering results using semantic structures," in *Proc. of the 7th European Conf. on Principles and Practice of Knowledge Discovery in Databases (PKDD-03)*, 2003.

[2] A. Rauber, "LabelSOM: on the labeling of self-organizing maps," in *Proc. of the 1999 Int'l Joint Conf. on Neural Networks (IJCNN-99)*, 1999, pp. 3527–3532.

[3] Q. Mei, X. Shen, and C. Zhai, "Automatic labeling of multinomial topic models," in *Proc. of the 13th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining (KDD-07)*, 2007, pp. 490–499.

[4] R. Agrawal, J. Gehrke, D. Gunopulos, and P. Raghavan, "Automatic subspace clustering of high dimensional data for data mining applications," in *Proc. of the 1998 ACM SIGMOD Int'l Conf. on Management of Data (SIGMOD-98)*, 1998, pp. 94–105.

[5] Y. Kameya, S. Nakamura, T. Iwasaki, and T. Sato, "Verbal characterization of probabilistic clusters by minimal discriminative propositions," Dept. of Computer Science, Tokyo Institute of Technology, Technical Report TR11-0001, 2011, http://www.cs.titech.ac.jp/~tr/ . Also available as arXiv:1108.5002, http://arxiv.org/abs/1108.5002 .

[6] P. Kralj Novak, N. Lavrač, and G. I. Webb, "Supervised descriptive rule discovery: a unifying survey of contrast set, emerging pattern and subgroup mining," *J. of Machine Learning Research*, vol. 10, pp. 377–403, 2009.

[7] L. Geng and H. J. Hamilton, "Interestingness measures for data mining: a survey," *ACM Computing Surveys*, vol. 38, no. 3, pp. 1–32, 2006.

[8] H. Fan and K. Ramamohanarao, "A Bayesian approach to use emerging patterns for classification," in *Proc. of the 14th Australasian Database Conf. (ADC-03)*, 2003, pp. 39–48.

[9] P. Cheeseman and J. Stutz, "Bayesian classification (AutoClass): theory and results," in *Advances in Knowledge Discovery and Data Mining*. The MIT Press, 1995.